日 本 国 特 許 庁
JAPAN PATENT OFFICE

別紙添付の書類に記載されている事項は下記の出願書類に記載されている事項と同一であることを証明する。

This is to certify that the annexed is a true copy of the following application as filed with this Office.

出 願 年 月 日
Date of Application: ２００２年１２月１７日

出 願 番 号
Application Number: 特願２００２－３６５７６４

パリ条約による外国への出願に用いる優先権の主張の基礎となる出願の国コードと出願番号
The country code and number of your priority application, to be used for filing abroad under the Paris Convention, is ＪＰ２００２－３６５７６４

出 願 人
Applicant(s): 独立行政法人科学技術振興機構

２０１１年 ４月１２日

特許庁長官
Commissioner,
Japan Patent Office 岩井良行

出証番号 出証特２０１１－３０１２９０３

【書類名】　特許願
【整理番号】　P2234JST
【特記事項】　特許法第３０条第１項の規定の適用を受けようとする特許出願
【提出日】　平成14年12月17日
【あて先】　特許庁長官　殿
【発明者】
　【住所又は居所】　千葉県佐倉市臼井８６
　【氏名】　中臺　一博
【発明者】
　【住所又は居所】　東京都渋谷区西原２－１０－９
　【氏名】　奥乃　博
【発明者】
　【住所又は居所】　埼玉県川越市西小仙波町２－１８－３
　【氏名】　北野　宏明
【特許出願人】
　【識別番号】　396020800
　【氏名又は名称】　科学技術振興事業団
【代理人】
　【識別番号】　100082876
　【弁理士】
　【氏名又は名称】　平山　一幸
　【電話番号】　03-3352-1808
【選任した代理人】
　【識別番号】　100069958
　【弁理士】
　【氏名又は名称】　海津　保三

【手数料の表示】

【予納台帳番号】　031727

【納付金額】　21,000円

【提出物件の目録】

【物件名】　明細書　1

【物件名】　図面　1

【物件名】　要約書　1

【包括委任状番号】　0013677

【プルーフの要否】　要

【書類名】　明細書

【発明の名称】　ロボット視聴覚システム

【特許請求の範囲】

【請求項１】　各話者が発した単語とその方向とを組み合わせて成る複数の音響モデルと、これらの音響モデルを使用して、音源分離された音響信号に対して音声認識プロセスを実行する音声認識エンジンと、前記音声認識プロセスによって前記音響モデル別に得られた複数の音声認識プロセス結果を統合し、何れかの音声認識プロセス結果を選択するセレクタと、を備え、

各話者が同時に発話した単語を各々認識することを特徴とする、ロボット視聴覚システム。

【請求項２】　前記セレクタが、多数決により前記音声認識プロセス結果を選択するように構成されていることを特徴とする、請求項１に記載のロボット視聴覚システム。

【請求項３】　前記セレクタにて選択された音声認識プロセス結果を外部に出力する対話部を備えていることを特徴とする、請求項１又は２に記載のロボット視聴覚システム。

【請求項４】　外部の音を集音する少なくとも一対のマイクを備えており、このマイクからの音響信号に基づいて、ピッチ抽出，調波構造に基づいたグルーピングによる音源の分離及び定位によって少なくとも一人の話者の方向を決定し、その聴覚イベントを抽出する聴覚モジュールと、

ロボットの前方を撮像するカメラを備えており、このカメラにより撮像された画像に基づいて各話者の顔識別と定位とから、各話者を同定してその顔イベントを抽出する顔モジュールと、

ロボットを水平方向に回動させる駆動モータを備えこの駆動モータの回転位置に基づいてモータイベントを抽出するモータ制御モジュールと、

上記聴覚イベント，顔イベント及びモータイベントから、聴覚イベントの音源定位及び顔イベントの顔定位の方向情報に基づいて各話者の方向を決定し、この決定に対してカルマンフィルタを用いて上記イベントを時間方向に接続することにより聴覚ストリーム及び顔ストリームを生成し、さらにこれらを関連付けてア

ソシエーションストリームを生成するアソシエーションモジュールと、

これらのストリームに基づいてアテンション制御と、それに伴う行動のプランニング結果に基づいて、モータの駆動制御を行うアテンション制御モジュールと、を備え、

上記聴覚モジュールが、

上記アソシエーションモジュールからの正確な音源方向情報に基づいて、正面方向で最小となり且つ左右に角度が大きくなるにつれて大きくなるパスレンジを有するアクティブ方向通過型フィルタにより、所定幅の範囲内の両耳間位相差（ＩＰＤ）または両耳間強度差（ＩＩＤ）をもったサブバンドを集めて、音源の波形を再構築することにより、音源分離を行なうと共に、

複数の音響モデルを使用して音源分離された音響信号の音声認識を行ない、各音響モデルによる音声認識結果をセレクタにより統合して、これらの音声認識結果のうち最も信頼性の高い音声認識結果を判断するように構成されていることを特徴とする、ロボット視聴覚システム。

【請求項５】　外部の音を集音する少なくとも一対のマイクを備えており、このマイクからの音響信号に基づいて、ピッチ抽出，調波構造に基づいたグルーピングによる音源の分離及び定位によって少なくとも一人の話者の方向を決定し、その聴覚イベントを抽出する聴覚モジュールと、

ロボットの前方を撮像するカメラを備えており、このカメラにより撮像された画像に基づいて各話者の顔識別と定位とから、各話者を同定してその顔イベントを抽出する顔モジュールと、

ステレオカメラにより撮像された画像から抽出された視差に基づいて縦に長い物体を抽出定位して、ステレオイベントを抽出するステレオモジュールと、

ロボットを水平方向に回動させる駆動モータを備えこの駆動モータの回転位置に基づいてモータイベントを抽出するモータ制御モジュールと、

上記聴覚イベント，顔イベント，ステレオイベント及びモータイベントから、聴覚イベントの音源定位及び顔イベントの顔定位の方向情報に基づいて各話者の方向を決定し、この決定に対してカルマンフィルタを用いて上記イベントを時間方向に接続することにより聴覚ストリーム，顔ストリーム及びステレオ視覚スト

リームを生成し、さらにこれらを関連付けてアソシエーションストリームを生成するアソシエーションモジュールと、

これらのストリームに基づいてアテンション制御と、それに伴う行動のプランニング結果に基づいて、モータの駆動制御を行うアテンション制御モジュールと、を備え、

上記聴覚モジュールが、

上記アソシエーションモジュールからの正確な音源方向情報に基づいて、正面方向で最小となり且つ左右に角度が大きくなるにつれて大きくなるパスレンジを有するアクティブ方向通過型フィルタにより、所定幅の範囲内の両耳間位相差（ＩＰＤ）または両耳間強度差（ＩＩＤ）をもったサブバンドを集めて、音源の波形を再構築することにより音源分離を行なうと共に、

複数の音響モデルを使用して音源分離された音響信号の音声認識を行ない、各音響モデルによる音声認識結果をセレクタにより統合して、これらの音声認識結果のうち最も信頼性の高い音声認識結果を判断するように構成されていることを特徴とする、ロボット視聴覚システム。

【請求項６】　前記聴覚モジュールによる音声認識ができなかったときに、前記アテンション制御モジュールが、当該音響信号の音源の方向に前記マイク及び前記カメラを向けて前記マイクから再び音声を集音させ、この音に対して前記聴覚モジュールにより音源定位・分離された音響信号に基づいて再度聴覚モジュールによる音声認識を行なうように構成されていることを特徴とする、請求項４又は５に記載のロボット視聴覚システム。

【請求項７】　前記聴覚モジュールが、音声認識を行なう際に、顔モジュールによる顔イベント又は／及びステレオモジュールによるステレオイベントを参照することを特徴とする、請求項５又は６に記載のロボット視聴覚システム。

【請求項８】　前記聴覚モジュールにて判断された音声認識結果を外部に出力する対話部を備えていることを特徴とする、請求項４～７の何れかに記載のロボット視聴覚システム。

【請求項９】　前記アクティブ方向通過型フィルタのパスレンジが、周波数毎に制御可能であることを特徴とする、請求項４～８の何れかに記載のロボット

視聴覚システム。

【発明の詳細な説明】

【０００１】

【発明の属する技術分野】

本発明はロボット、特に人型または動物型ロボットにおける視聴覚システムに関するものである。

【０００２】

【従来の技術】

近年、このような人型または動物型ロボットにおいては、ＡＩの研究目的の対象にとどまらず、所謂「人間のパートナー」としての将来的な利用が考えられている。そして、ロボットが人間との知的なソーシャルインタラクションを行なうために、視聴覚等の知覚がロボットには必要である。そして、ロボットが人間とのソーシャルインタラクションを実現するためには、知覚のうち、視聴覚、特に聴覚が重要な機能であることは明らかである。従って、視覚，聴覚に関して、所謂能動知覚が注目されてきている。

【０００３】

ここで、能動知覚とは、ロボット視覚やロボット聴覚等の知覚を担当する知覚装置を知覚すべき目標に追従する働きをを言い、例えば、これらの知覚装置を支持する頭部を駆動機構により目標に追従するように姿勢制御するものである。

【０００４】

ロボットにおける能動視覚においては、少なくとも知覚装置であるカメラが、駆動機構による姿勢制御によって、その光軸方向を目標に向かって保持され、更に目標に対して自動的にフォーカシングやズームイン，ズームアウト等を行う。これにより、目標が移動してもカメラによって撮像される。このような能動視覚の研究が従来、様々に行なわれている。

【０００５】

これに対して、ロボットにおける能動聴覚においては、少なくとも知覚装置であるマイクが、駆動機構による姿勢制御によってその指向性を目標に向かって保持され、目標からの音がマイクによって集音される。このとき、能動聴覚の不利

な点として、駆動機構が作用している間はマイクが駆動機構の作動音を拾ってしまうために目標からの音に比較的大きなノイズが混入してしまい、目標からの音を認識できなくなってしまうことがある。このような能動聴覚の不利な点を排除するために、例えば視覚情報を参照して音源の方向付けを行なうことにより、目標からの音を正確に認識する方法が採用されている。

【０００６】

ところで、このような能動聴覚においては、マイクで集音した音に基づいて、（Ａ）音源の定位，（Ｂ）各音源から発せられた音毎の分離，（Ｃ）そして各音源からの音の認識を行なう必要がある。このうち、（Ａ）音源定位及び（Ｂ）音源分離については、能動聴覚における実時間・実環境での音源定位・追跡・分離に関する種々の研究が行なわれている（特許文献１参照）。

【０００７】

【特許文献１】

国際公開第０１／９５３１４号パンフレット

【０００８】

ここで、例えば、特許文献１に示すように、ＨＲＴＦ（頭部伝達関数）から求められる両耳間位相差（ＩＰＤ），両耳間強度差（ＩＩＤ）を利用して音源定位を行なうことが知られている。また、特許文献１では、例えば所謂方向通過型フィルタ、即ちディレクションパスフィルタを用いて、特定の方向のＩＰＤと同じＩＰＤを有するサブバンドを選択することにより、各音源からの音を分離する方法が知られている。

【０００９】

これに対して、音源分離により分離された各音源からの音の認識については、例えばマルチコンディショニングやミッシングデータ等のノイズに対してロバストな音声認識へのアプローチは種々の研究が行なわれている（例えば非特許文献１，２参照）。

【００１０】

【非特許文献１】

Ｊ．ベーカー等著，クリーンスピーチモデルに基づくロバスト“ユー

ロスピーチ２００１－第７回ヨーロッパ会議予稿集”，２００１年，第１巻，p２１３－２１６(J.Baker, M.Cooke, and P.Green, Robust as based on clean speechmodels: An evaluation of missing data techniques for connected digit recognition in noise. ”7th European conference on Speech Commnication Technology”, Volume 1, p. 213-216)

【非特許文献２】

Ｐ．レネベイ等著，ロバストスピーチ認識 ”ユーロスピーチ２００１－第７回ヨーロッパ会議予稿集”，２００１年，第１２巻，pp.１１０７－１１１０ (Philippe Renevey, Rolf Vetter, and Jens Kraus. Robust speech recognition using missing feature theory and vector quantization. ”7th European Conference on Speech Communication Technology”, Volume 12, pp. 1107-1110)

【００１１】

【発明が解決しようとする課題】

しかしながら、これらの研究（例えば非特許文献１，２）においては、Ｓ／Ｎ比が小さい場合には、有効な音声認識を行なうことができない。また、実時間・実環境での音声認識についての研究は行なわれていない。

【００１２】

この発明は、以上の点に鑑みて、各音源からの分離された音についての認識を行なうようにしたロボット視聴覚システムを提供することを目的としている。

【００１３】

【課題を解決するための手段】

上記目的を達成するために、本発明のロボット視聴覚システムの第１の構成は、各話者が発した単語とその方向とを組み合わせて成る複数の音響モデルと、これらの音響モデルを使用して、音源分離された音響信号に対して音声認識プロセスを実行する音声認識エンジンと、この音声認識プロセスによって音響モデル別に得られた複数の音声認識プロセス結果を統合し、何れかの音声認識プロセス結果を選択するセレクタと、を備えて、各話者が同時に発話した単語を各々認識することを特徴としている。

【００１４】

前記セレクタは、多数決により前記音声認識プロセス結果を選択するように構成され、前記セレクタにて選択された音声認識プロセス結果を外部に出力する対話部を備えていてもよい。

【００１５】

このような第１の構成によれば、音源定位・音源分離された音響信号に基づいて、複数の音響モデルを使用することによって、それぞれ音声認識プロセスを行なう。そして、各音響モデルによる音声認識プロセス結果をセレクタにより統合して、最も信頼性の高い音声認識結果を判断する。

【００１６】

また、上記目的を達成するために、本発明のロボット視聴覚システムの第２の構成は、外部の音を集音する少なくとも一対のマイクを備えており、このマイクからの音響信号に基づいて、ピッチ抽出，調波構造に基づいたグルーピングによる音源の分離及び定位によって少なくとも一人の話者の方向を決定し、その聴覚イベントを抽出する聴覚モジュールと、ロボットの前方を撮像するカメラを備えており、このカメラにより撮像された画像に基づいて各話者の顔識別と定位とから各話者を同定してその顔イベントを抽出する顔モジュールと、ロボットを水平方向に回動させる駆動モータを備えこの駆動モータの回転位置に基づいてモータイベントを抽出するモータ制御モジュールと、上記聴覚イベント，顔イベント及びモータイベントから、聴覚イベントの音源定位及び顔イベントの顔定位の方向情報に基づいて各話者の方向を決定し、この決定に対してカルマンフィルタを用いて上記イベントを時間方向に接続することにより聴覚ストリーム及び顔ストリームを生成し、さらにこれらを関連付けてアソシエーションストリームを生成するアソシエーションモジュールと、これらのストリームに基づいてアテンション制御と、それに伴う行動のプランニング結果に基づいてモータの駆動制御を行うアテンション制御モジュールと、を備え、上記聴覚モジュールが、上記アソシエーションモジュールからの正確な音源方向情報に基づいて、正面方向で最小となり且つ左右に角度が大きくなるにつれて大きくなるパスレンジを有するアクティブ方向通過型フィルタにより、所定幅の範囲内の両耳間位相差（ＩＰＤ）または

両耳間強度差（ＩＩＤ）をもったサブバンドを集めて、音源の波形を再構築することにより音源分離を行なうと共に、複数の音響モデルを使用して音源分離された音響信号の音声認識を行ない、各音響モデルによる音声認識結果をセレクタにより統合して、これらの音声認識結果のうち最も信頼性の高い音声認識結果を判断するように構成されている。

【００１７】

このような第２の構成によれば、聴覚モジュールがマイクが集音した外部の対象からの音から、調波構造を利用してピッチ抽出を行なうことにより音源毎の方向を得て個々の話者を同定し、その聴覚イベントを抽出する。

【００１８】

また、顔モジュールが、カメラにより撮像された画像から、パターン認識による各話者の顔識別と定位から、個々の話者の顔イベントを抽出する。

【００１９】

さらに、モータ制御モジュールが、ロボットを水平方向に回動させる駆動モータの回転位置に基づいて、ロボットの方向を検出することによって、モータイベントを抽出する。

【００２０】

なお、上記イベントとは、各時点において検出される音または顔が在ること、あるいは駆動モータが回転される状態を示しており、ストリームとは、エラー訂正処理を行ないながら例えばカルマンフィルタ等により時間的に連続するように接続したイベントを示している。

【００２１】

ここで、アソシエーションモジュールは、このようにしてそれぞれ抽出された聴覚イベント，顔イベント及びモータイベントに基づいて、各話者の聴覚ストリーム及び顔ストリームを生成し、さらにこれらのストリームを関連付けてアソシエーションストリームを生成して、アテンション制御モジュールがこれらのストリームに基づいてアテンション制御を行なうことにより、モータ制御モジュールの駆動モータ制御のプランニングを行なう。なお、アソシエーションストリームとは、聴覚ストリーム及び顔ストリームを包含する概念である。

【００２２】

なお、アテンションとは、ロボットが対象である話者を聴覚的及び／または視覚的に「注目」することであり、アンテンション制御とは、モータ制御モジュールによりその向きを変えることによってロボットが上記話者に注目するようにすることである。

【００２３】

そして、アテンション制御モジュールは、このプランニングに基づいて、モータ制御モジュールの駆動モータを制御することにより、ロボットの方向を対象である話者に向ける。これにより、ロボットが対象である話者に対して正対することにより、聴覚モジュールが当該話者の声を感度の高い正面方向にてマイクで正確に集音，定位することができると共に、顔モジュールが当該話者の画像をカメラにより良好に撮像することができるようになる。

【００２４】

従って、このような聴覚モジュール，顔モジュール及びモータ制御モジュールと、アソシエーションモジュール及びアテンション制御モジュールとの連携によって、ロボットの聴覚及び視覚がそれぞれ有する曖昧性が互いに補完されることになり、所謂ロバスト性が向上し、複数の話者であっても、各話者をそれぞれ知覚することができる。

【００２５】

また、例えば聴覚イベントまたは顔イベントの何れか一方が欠落したときであっても、顔イベントまたは聴覚イベントのみに基づいて、対象である話者をアソシエーションモジュールが知覚することができるので、リアルタイムにモータ制御モジュールの制御を行なうことができる。

【００２６】

さらに、上記聴覚モジュールが、上述したように音源定位・音源分離された音響信号に基づいて、複数の音響モデルを使用することによってそれぞれ音声認識を行なう。そして、各音響モデルによる音声認識結果をセレクタにより統合して、最も信頼性の高い音声認識結果を判断する。

【００２７】

これにより、従来の音声認識と比較して複数の音響モデルを使用することによって、実時間・実環境での正確な音声認識を行なうことが可能になると共に、各音響モデルによる音声認識結果をセレクタにより統合して、最も信頼性の高い音声認識結果を判断して、より一層正確な音声認識を行なうことができる。

【００２８】

また、上記目的を達成するために、本発明のロボット視聴覚システムの第３の構成は、外部の音を集音する少なくとも一対のマイクを備えており、このマイクからの音響信号に基づいてピッチ抽出，調波構造に基づいたグルーピングによる音源の分離及び定位によって少なくとも一人の話者の方向を決定しその聴覚イベントを抽出する聴覚モジュールと、ロボットの前方を撮像するカメラを備えこのカメラで撮像された画像に基づいて各話者の顔識別と定位とから各話者を同定してその顔イベントを抽出する顔モジュールと、ステレオカメラにより撮像された画像から抽出された視差に基づいて縦に長い物体を抽出定位してステレオイベントを抽出するステレオモジュールと、ロボットを水平方向に回動させる駆動モータを備えこの駆動モータの回転位置に基づいてモータイベントを抽出するモータ制御モジュールと、前記聴覚イベント，顔イベント，ステレオイベント及びモータイベントから聴覚イベントの音源定位及び顔イベントの顔定位の方向情報に基づいて各話者の方向を決定しこの決定に対してカルマンフィルタを用いて前記イベントを時間方向に接続することにより聴覚ストリーム，顔ストリーム及びステレオ視覚ストリームを生成しさらにこれらを関連付けてアソシエーションストリームを生成するアソシエーションモジュールと、これらのストリームに基づいてアテンション制御と、それに伴う行動のプランニング結果に基づいてモータの駆動制御を行うアテンション制御モジュールと、を備え、上記聴覚モジュールが、上記アソシエーションモジュールからの正確な音源方向情報に基づいて、正面方向で最小となり且つ左右に角度が大きくなるにつれて大きくなるパスレンジを有するアクティブ方向通過型フィルタにより、所定幅の範囲内の両耳間位相差（ＩＰＤ）または両耳間強度差（ＩＩＤ）をもったサブバンドを集めて、音源の波形を再構築することにより音源分離を行なうと共に、音声認識の際に、複数の音響モデルを使用して音源分離された音響信号の音声認識を行ない、各音響モデルに

よる音声認識結果をセレクタにより統合して、これらの音声認識結果のうち最も信頼性の高い音声認識結果を判断するように構成されている。

【００２９】

このような第３の構成によれば、聴覚モジュールは、マイクが集音した外部の目標からの音から調波構造を利用してピッチ抽出を行なうことにより音源毎の方向を得て、個々の話者の方向を決定してその聴覚イベントを抽出する。

【００３０】

また、顔モジュールは、カメラにより撮像された画像からパターン認識による各話者の顔識別と定位から各話者を同定して、個々の話者の顔イベントを抽出する。さらに、ステレオモジュールは、ステレオカメラにより撮像された画像から抽出された視差に基づいて縦に長い物体を抽出定位してステレオイベントを抽出する。

【００３１】

さらに、モータ制御モジュールは、ロボットを水平方向に回動させる駆動モータの回転位置に基づいて、ロボットの方向を検出することによってモータイベントを抽出する。

【００３２】

なお、上記イベントとは、各時点において検出される音，顔及び縦に長い物体が在ること、あるいは駆動モータが回転される状態を示しており、ストリームとは、エラー訂正処理を行ないながら例えばカルマンフィルタ等により時間的に連続するように接続したイベントを示している。

【００３３】

ここで、アソシエーションモジュールは、このようにしてそれぞれ抽出された聴覚イベント，顔イベント，ステレオイベント及びモータイベントに基づいて、聴覚イベントの音源定位及び顔イベントの顔定位の方向情報によって各話者の方向を決定することにより、各話者の聴覚ストリーム，顔ストリーム及びステレオ視覚ストリームを生成し、さらにこれらのストリームを関連付けてアソシエーションストリームを生成する。なお、アソシエーションストリームとは、聴覚ストリーム，顔ストリーム及びステレオ視覚ストリームを包含する概念である。この

際、アソシエーションモジュールは、聴覚イベントの音源定位及び顔イベントの顔定位、即ち聴覚及び視覚の方向情報に基づいて各話者の方向を決定し、決定された各話者の方向を参考にして、アソシエーションストリームを生成する。

【００３４】

そして、アテンション制御モジュールが、これらのストリームに基づいてアテンション制御と、それに伴う行動のプランニング結果に基づいて、モータの駆動制御を行なう。そして、アテンション制御モジュールは、このプランニングに基づいてモータ制御モジュールの駆動モータを制御してロボットの方向を目標である話者に向ける。これにより、ロボットが目標である話者に対して正対することによって聴覚モジュールが当該話者の声を感度の高い正面方向にてマイクにより正確に集音，定位することができる共に、顔モジュールが当該話者の画像をカメラにより良好に撮像することができるようになる。

【００３５】

従って、このような聴覚モジュール，顔モジュール，ステレオモジュール及びモータ制御モジュールと、アソシエーションモジュール及びアテンション制御モジュールとの連携によって、聴覚ストリームの音源定位及び顔ストリームの話者定位という方向情報に基づいて各話者の方向を決定することにより、ロボットの聴覚及び視覚がそれぞれ有する曖昧性が互いに補完されることになり、所謂ロバスト性が向上し、複数の話者であっても各話者をそれぞれ確実に知覚することができる。

【００３６】

また、例えば聴覚ストリーム，顔ストリーム及びステレオ視覚ストリームの何れかが欠落したときであっても、残りのストリームに基づいて目標である話者をアテンション制御モジュールが追跡することができるので、正確に目標の方向を把握して、モータ制御モジュールの制御を行なうことができる。

【００３７】

ここで、聴覚モジュールが、アソシエーションモジュールからのアソシエーションストリームを参照することにより、顔モジュールからの顔ストリームやステレオモジュールからのステレオ視覚ストリームをも考慮して音源定位を行なうこ

とによって、より一層正確な音源定位を行なうことができる。

【００３８】

そして、上記聴覚モジュールは、アソシエーションモジュールからの正確な音源方向情報に基づいて、聴覚特性に従って正面方向で最小となり且つ左右に角度が大きくなるにつれて大きくなるパスレンジを有するアクティブ方向通過型フィルタにより、所定幅の範囲内の両耳間位相差（ＩＰＤ）または両耳間強度差（ＩＤ）をもったサブバンドを集めて、音源の波形を再構築して音源分離を行なうので、上述した聴覚特性に応じてパスレンジ即ち感度を調整することにより、方向による感度の違いを考慮して、より正確に音源分離を行なうことができる。さらに、上記聴覚モジュールは、上述したように聴覚モジュールによって音源定位・音源分離された音響信号に基づいて、複数の音響モデルを使用することによってそれぞれ音声認識を行なう。そして、各音響モデルによる音声認識結果をセレクタにより統合して、最も信頼性の高い音声認識結果を判断して、この音声認識結果を対応する話者と関連付けて出力する。

【００３９】

これにより、従来の音声認識と比較して、複数の音響モデルを使用することによって、実時間・実環境での正確な音声認識を行なうことが可能になると共に、各音響モデルによる音声認識結果をセレクタにより統合して、最も信頼性の高い音声認識結果を判断することにより、より一層正確な音声認識を行なうことができる。

【００４０】

なお、第２の構成と第３の構成においては、聴覚モジュールによる音声認識ができなかったときに、前記アテンション制御モジュールが、当該音響信号の音源の方向に前記マイク及び前記カメラを向けて、前記マイクから再び音声を集音させ、この音に対して聴覚モジュールにより音源定位・分離された音響信号に基づいて、再度聴覚モジュールによる音声認識を行なうように構成されている。

【００４１】

さらに、前記聴覚モジュールは、音声認識を行なう際に顔モジュールによる顔イベントを参照するのが望ましい。また、前記聴覚モジュールにて判断された音

声認識結果を外部に出力する対話部が備えられていてもよい。さらに、前記アクティブ方向通過型フィルタのパスレンジが周波数毎に制御可能であることが望ましい。

【００４２】

上記聴覚モジュールによる音声認識ができなかったとき、アテンション制御モジュールが、当該音響信号の音源の方向（当該話者）にマイク及びカメラを向けて、再度マイクから音声を集音させ、聴覚モジュールにより音源定位・分離された音響信号に基づいて、再度聴覚モジュールによる音声認識を行なう場合には、ロボットの聴覚モジュールのマイク及び顔モジュールのカメラが当該話者と正対することによって、確実な音声認識を行なうことが可能になる。

【００４３】

上記聴覚モジュールは、音声認識を行なう際に、アソシエーションモジュールからのアソシエーションストリームを参照することにより、顔モジュールからの顔ストリームをも考慮する。即ち、聴覚モジュールは、顔モジュールにより定位された顔イベントに関して、聴覚モジュールにより定位・分離された音源（話者）からの音響信号に基づいて音声認識を行なうことにより、より一層正確な音声認識を行なうことができる。

【００４４】

上記アクティブ方向通過型フィルタのパスレンジが周波数毎に制御可能であると、さらに集音した音からの分離の精度が上がり、これにより音声認識もさらに向上する。

【００４５】

【発明の実施の形態】

以下、図面に示した実施形態に基づいて、この発明を詳細に説明する。

図１及び図２は、それぞれこの発明によるロボット視聴覚システムの一実施形態を備えた実験用の上半身のみの人型ロボットの全体構成例を示している。図１において、人型ロボット１０は、４ＤＯＦ（自由度）のロボットとして構成されており、ベース１１と、ベース１１上にて一軸（垂直軸）周りに回動可能に支持された胴体部１２と、胴体部１２上にて三軸方向（垂直軸，左右方向の水平軸及

び前後方向の水平軸）の周りに揺動可能に支持された頭部１３とを含んでいる。

【００４６】

上記ベース１１は固定配置されていてもよく、脚部として動作可能としてもよい。また、ベース１１は、移動可能な台車等の上に載置されていてもよい。胴体部１２は、ベース１１に対して垂直軸の周りに、図１にて矢印Ａで示すように回動可能に支持されており、図示しない駆動手段によって回転駆動されると共に、図示の場合、防音性の外装によって覆われている。

【００４７】

頭部１３は胴体部１２に対して連結部材１３ａを介して支持されており、この連結部材１３ａに対して前後方向の水平軸の周りに、図１にて矢印Ｂで示すように揺動可能に、また左右方向の水平軸の周りに、図２にて矢印Ｃで示すように揺動可能に支持されていると共に、上記連結部材１３ａが、胴体部１２に対してさらに前後方向の水平軸の周りに、図１にて矢印Ｄで示すように揺動可能に支持されており、それぞれ図示しない駆動手段によって、各矢印Ａ，Ｂ，Ｃ，Ｄ方向に回転駆動される。ここで、頭部１３は、図３に示すように全体が防音性の外装１４により覆われ、前側にロボット視覚を担当する視覚装置としてのカメラ１５、両側にロボット聴覚を担当する聴覚装置としての一対のマイク１６（１６ａ，１６ｂ）を備えている。なお、マイク１６は、頭部１３の両側に限定されることなく、頭部１３の他の位置あるいは胴体部１２等に設けられていてもよい。

【００４８】

上記外装１４は、例えばウレタン樹脂等の吸音性の合成樹脂から構成されており、頭部１３の内部がほぼ完全に密閉されて、頭部１３の内部の遮音が行われるように構成されている。なお、胴体部１２の外装も、頭部１３の外装１４と同様に、吸音性の合成樹脂から構成されている。

【００４９】

上記カメラ１５は公知の構成であって、例えば所謂パン，チルト，ズームの３ＤＯＦ（自由度）を有する市販のカメラにより構成されている。尚、カメラ１５は、同期をとってステレオ画像を送ることができるように設計されている。

【００５０】

上記マイク１６は、それぞれ頭部１３の側面において前方に向かって指向性を有するように取り付けられている。マイク１６の左右の各マイク１６ａ，１６ｂは、それぞれ図１及び図２に示すように、頭部１３の外装１４の両側に配置された段部１４ａ，１４ｂの内側に取り付けられている。そして、各マイク１６ａ，１６ｂは、段部１４ａ，１４ｂに設けられた貫通穴を通して、前方の音を集音すると共に、外装１４の内部の音を拾わないように、適宜の手段により遮音されている。なお、段部１４ａ，１４ｂに設けられた貫通穴は、段部１４ａ，１４ｂの内側から頭部前方に向けて貫通するように、各段部１４ａ，１４ｂに形成されている。これにより、各マイク１６ａ，１６ｂは、所謂バイノーラルマイクとして構成されている。なお、マイク１６ａ，１６ｂの取付位置に近接する外装１４は人間の外耳形状に形成されていてもよい。ここで、マイク１６は、外装１４の内側に配置された一対の内部マイクを含んでいてもよく、この内部マイクにより集音された内部音に基づいて、ロボット１０の内部に発生するノイズをキャンセルすることができる。

【００５１】

図４は、上記カメラ１５及びマイク１６を含むロボット視聴覚の電気的構成例を示している。図４において、ロボット視聴覚システム１７は、聴覚モジュール２０，顔モジュール３０，ステレオモジュール３７，モータ制御モジュール４０及びアソシエーションモジュール５０から構成されている。

【００５２】

ここで、アソシエーションモジュール５０はクライアントからの依頼に応じて処理を実行するサーバとして構成されており、このサーバに対するクライアントが、他のモジュール、即ち聴覚モジュール２０，顔モジュール３０，ステレオモジュール３７，モータ制御モジュール４０であり、これらのサーバとクライアントとは、互いに非同期で動作する。なお、上記サーバと各クライアントとは、各々、パーソナルコンピュータにより構成されており、更にこれらの各パーソナルコンピュータは、例えばＴＣＰ／ＩＰプロトコルの通信環境の下で、相互にＬＡＮ(Local Area Network)として構成されている。この場合、好ましくは、データ量の大きいイベントやストリームの通信のためには、ギガビット(Giga bit)のデ

ータ交換が可能な高速ネットワークをロボット視聴覚システム１７に適用するのが好ましく、また時刻の同期等の制御用通信のためには中速ネットワークをロボット視聴覚システム１７に適用するのが好ましい。このように大きなデータが高速に各パーソナルコンピュータ間を伝送することで、ロボット全体のリアルタイム性及びスケーラビリティを向上させることができる。

【００５３】

また、各モジュール２０，３０，３７，４０，５０は、それぞれ階層的に分散して構成されており、具体的には下位から順次にデバイス層，プロセス層，特徴層，イベント層から構成されている。

【００５４】

上記聴覚モジュール２０は、デバイス層としてのマイク１６と、プロセス層としてのピーク抽出部２１，音源定位部２２，音源分離部２３及びアクティブ方向通過型フィルタ２３ａと、特徴層（データ）としてのピッチ２４，音源水平方向２５と、イベント層としての聴覚イベント生成部２６と、さらにプロセス層としての音声認識部２７及び会話部２８と、から構成されている。

【００５５】

ここで、聴覚モジュール２０は、図５に示すように作用する。即ち、図５において、聴覚モジュール２０は、例えば４８ｋＨｚ，１６ビットでサンプリングされたマイク１６からの音響信号を、符号Ｘ１で示すようにＦＦＴ（高速フーリエ変換）により周波数解析して、符号Ｘ２で示すように左右のチャンネル毎にスペクトルを生成する。そして、聴覚モジュール２０は、ピーク抽出部２１により左右のチャンネル毎に一連のピークを抽出して、左右のチャンネルで同じか類似のピークをペアとする。

【００５６】

ここで、ピーク抽出は、（α）パワーがしきい値以上で且つ（β）ローカルピークであって、（γ）低周波ノイズとパワーの小さい高周波帯域をカットするため例えば９０Ｈｚ乃至３ｋＨｚの間の周波数であるという、３つの条件（α～γ）を満たすデータのみを透過させる帯域フィルタを使用して行なわれる。このしきい値は、周囲の暗騒音を計測して、さらに感度パラメータ、例えば１０ｄＢを

加えた値として定義される。

【００５７】

そして、聴覚モジュール２０は、各ピークが調波構造を有していることを利用して音源分離を行う。具体的には、音源分離部２３は、周波数の低い方から順に調波構造を有するローカルピークを抽出して、この抽出されたピークの集合を一つの音とみなす。このようにして、音源毎の音響信号が混合音からそれぞれ分離される。音源分離の際、聴覚モジュール２０の音源定位部２２は、符号Ｘ３で示すように、音源分離部２３にて分離された各音源毎の音響信号に対して、左右のチャンネルから同じ周波数の音響信号を選択して、ＩＰＤ（相互位相差）及びＩＤ（相互強度差）を計算する。なお、この計算は、例えば５度毎に行われる。そして、音源定位部２２は、計算結果をアクティブ方向通過型フィルタ２３ａに出力する。

【００５８】

これに対して、アクティブ方向通過型フィルタ２３ａは、アソシエーションモジュール５０にて算出されたアソシエーションストリーム５９の方向θに基づいて、符号Ｘ４で示すように、ＩＰＤの理論値ＩＰＤ（$=\Delta\phi'(\theta)$）を生成すると共に、ＩＩＤの理論値ＩＩＤ（$=\Delta\rho'(\theta)$）を計算する。なお、方向θは、顔定位（顔イベント３９）とステレオ視覚（ステレオ視覚イベント３９ａ）と音源定位（聴覚イベント２９）とに基づいて、アソシエーションモジュール５０におけるリアルタイムトラッキング（符号Ｘ３′）による算出結果である。

【００５９】

ここで、理論値ＩＰＤと理論値ＩＩＤの各計算は、以下に説明する聴覚エピポーラ幾何を利用して行われ、具体的にはロボット１０の正面を０度と設定し、±９０度の範囲で理論値ＩＰＤ及び理論値ＩＩＤが計算される。ここで、上記聴覚エピポーラ幾何は、ＨＲＴＦを使用せずに音源の方向情報を得るために必要である。ステレオ視覚研究においては、エピポーラ幾何が最も一般的な定位法の一つであり、聴覚エピポーラ幾何は視覚におけるエピポーラ幾何の聴覚への応用である。そして、聴覚エピポーラ幾何が幾何学的関係を利用して方向情報を得るので、ＨＲＴＦを不要にすることができるのである。

【００６０】

上記聴覚エピポーラ幾何においては、音源が無限遠にあると仮定し、Δφ，θ，ｆ，ｖをそれぞれＩＰＤ，音源方向，周波数，音速とし、ｒをロボット頭部を球形とみなした場合の半径とすると、以下の式（１）

【数１】

$$\Delta\phi = \frac{2\pi f}{v} \times r(\theta + \sin\theta)$$

により表わされる。

【００６１】

他方、ＦＦＴ（高速フーリエ変換）により得られた一対のスペクトルに基づいて、各サブバンドのＩＰＤΔφ′及びＩＩＤΔρ′を、以下の式（２），（３）により計算する。

【数２】

$$\Delta\varphi' = \arctan\left(\frac{\Im[Sp_l]}{\Re[Sp_l]}\right) - \arctan\left(\frac{\Im[Sp_r]}{\Re[Sp_r]}\right)$$

【数３】

$$\Delta\rho' = 20\log_{10}\left(\frac{|Sp_l|}{|Sp_r|}\right)$$

ここで、Ｓｐｌ，Ｓｐｒは、それぞれある時刻に左右のマイク１６ａ，１６ｂから得られたスペクトルである。

【００６２】

さらに、アクティブ方向通過型フィルタ２３ａは、符号Ｘ７で示す通過帯域関数に従って、前記ストリーム方向θ_Sから、θ_Sに対応するアクティブ方向通過型フィルタ２３ａの通過帯域δ（θ_S）を選択する。ここで、通過帯域関数は、図５のＸ７に示すように、ロボットの正面方向（θ＝０度）で感度が最大となり、側方で感度が低下することから、θ＝０度で最小値をとり、側方でより大きくなるような関数である。これは、正面方向で定位の感度が最大になり、左右に角度が大きくなるにつれて感度が低下するという聴覚特性を再現するためのものである。なお、正面方向で定位の感度が最大になることは、哺乳類の目の構造に見

られる中心窩にならって聴覚中心窩と呼ぶ。この聴覚中心窩に関して、人間の場合には、正面の定位の感度が±２度程度であり、左右９０度付近にて±８度程度とされている。

【００６３】

そして、アクティブ方向通過型フィルタ２３ａは、選択した通過帯域δ（θ_S）を使用して、θ_L から θ_H の範囲にある音響信号を抽出する。尚、$\theta_L = \theta_S - \delta(\theta_S)$，$\theta_H = \theta_S + \delta(\theta_S)$ と定義する。

【００６４】

また、アクティブ方向通過型フィルタ２３ａは、符号Ｘ５で示すように、ストリーム方向 θ_S を頭部伝達関数（ＨＲＴＦ）に利用して、θ_L 及び θ_H におけるＩＰＤ及びＩＩＤの理論値ＩＰＤ（$=\Delta\phi_H(\theta_S)$）とＩＩＤ（$=\Delta\rho_H(\theta_S)$）とを、即ち抽出すべき音源の方向を推定する。そして、アクティブ方向通過型フィルタ２３ａは、音源方向θに対して聴覚エピポーラ幾何に基づいて各サブバンド毎に計算されたＩＰＤ（$=\Delta\phi_E(\theta)$）及びＩＩＤ（$=\Delta\rho_E(\theta)$）と、ＨＲＴＦに基づいて得られたＩＰＤ（$=\Delta\phi_H(\theta)$）及びＩＩＤ（$=\Delta\rho_H(\theta)$）とに基づいて、符号Ｘ６で示すように、前述した通過帯域δ（θ）により決定される角度 θ_L から θ_H の角度範囲で、抽出されたＩＰＤ（$=\Delta\phi_E$）及びＩＩＤ（$=\Delta\rho_E$）が以下の条件を満たすようなサブバンドを集める。

【００６５】

ここで、周波数 f_{th}は、フィルタリングの判断基準としてＩＰＤまたはＩＩＤを採用する閾値であって、ＩＰＤによる定位が有効である周波数の上限を示す。なお、周波数 f_{th}は、ロボット１０のマイク間距離に依存し、本実施形態においては例えば１５００Ｈｚ程度である。

【００６６】

即ち、

【数４】

$$f < f_{th} \quad : \quad \Delta\varphi_E(\theta_l) \le \Delta\varphi' \le \Delta\varphi_E(\theta_h)$$
$$f \ge f_{th} \quad : \quad \Delta\rho_H(\theta_l) \le \Delta\rho' \le \Delta\rho_H(\theta_h)$$

【００６７】

これは、所定周波数 f_{th}未満の周波数で、ＨＲＴＦによるＩＰＤの通過帯域δ（θ）の範囲内にＩＰＤ（＝Δφ′）が在る場合、そして所定周波数 f_{th}以上の周波数でＨＲＴＦによるＩＩＤの通過帯域δ（θ）の範囲内にＩＩＤ（＝Δρ′）が在る場合に、サブバンドを集めることを意味している。ここで、一般に低周波数帯域ではＩＰＤが大きく影響し、高周波数帯域ではＩＩＤが大きく影響し、その閾値である周波数 f_{th}はマイク間距離に依存する。

【００６８】

そして、アクティブ方向通過型フィルタ２３ａは、このようにして集めたサブバンドから音響信号を再合成して、波形を構築することにより、符号Ｘ８で示すように、パスーサブバンド方向を生成し、符号Ｘ９で示すように、各サブバンド毎にフィルタリングを行なって、符号Ｘ１０で示す逆周波数変換ＩＦＦＴ（逆フーリエ変換）により、符号Ｘ１１で示すように、該当範囲にある各音源からの分離音（音響信号）を抽出する。

【００６９】

上記音声認識部２７は、図５に示すように、自声抑制部２７ａと自動認識部２７ｂとから構成されている。自声抑制部２７ａは、聴覚モジュール２０にて音源定位・音源分離された各音響信号から、後述する対話部２８のスピーカ２８ｃから発せられた音声を除去して外部からの音響信号のみを取り出すものである。自動認識部２７ｂは、図６に示すように、音声認識エンジン２７ｃと音響モデル２７ｄとセレクタ２７ｅとから構成されており、この音声認識エンジン２７ｃとしては、例えば京都大学で開発された「Ｊｕｒｉａｎ」という音声認識エンジンを利用することができ、これにより各話者が発話した単語を認識することができるようになっている。

【００７０】

図６において、自動認識部２７ｂは、例として男性２人（話者Ａ，Ｃ）と女性１人（話者Ｂ）の三人の話者の認識を行なうように構成されている。このために自動認識部２７ｂには、各話者の各方向毎にそれぞれ音響モデル２７ｄが備えられている。図６の場合には、音響モデル２７ｄは、３人の各話者Ａ，Ｂ，Ｃに関してそれぞれ各話者が発した音声とその方向とを組み合わせて成り、複数種類、

この場合９種類の音響モデル２７ｄが備えられている。

【００７１】

音声認識エンジン２７ｃは、並列に９つの音声認識プロセスを実行し、その際に上記９つの音響モデル２７ｄが用いられる。具体的には、音声認識エンジン２７ｃは、それぞれ互いに並列的に入力された音響信号に対して、上記９つの音響モデル２７ｄを用いて音声認識プロセスを実行する。そして、これらの音声認識結果がセレクタ２７ｅに出力される。

【００７２】

上記セレクタ２７ｅは、各音響モデル２７ｄからのすべての音声認識プロセス結果を統合して、例えば多数決により最も信頼性が高い音声認識プロセス結果を判断して、その音声認識結果を出力する。

【００７３】

ここで、特定話者の音響モデル２７ｄに対する単語認識率を具体的な実験により説明する。

先ず、３ｍ×３ｍの部屋内において、３つのスピーカをロボット１０から１ｍの位置に且つロボットから０度及び±６０度の方向に置く。次に、音響モデル用の音声データとして、男性２名，女性１名が各々発話した、色，数字，食べ物のような１５０語の単語の音声をスピーカから出力して、ロボット１０のマイク１６ａ，１６ｂで集音する。なお、各単語の集音に当たり、一つのスピーカのみからの音声、二つのスピーカから同時に出力される音声、そして三つのスピーカから同時出力される音声、として、各単語に対して３つのパターンを録音する。そして、録音した音声信号に対して前述したアクティブ方向通過型フィルタ２３ａによって音声分離して各音声データを抽出し、話者及び方向毎に整理して、音響モデルのトレーニングセットを作成する。

【００７４】

そして、各音響モデル２７ｄには、トライフォンを使用して、各トレーニングセット毎に、ＨＴＫ（Ｈｉｄｄｅｎ　Ｍａｒｃｏｖ　Ｍｏｄｅｌ）ツールキット２７ｆを使用して、各話者の各方向毎に計９種類の音声認識用の音声データを作成した。

【００７５】

このようにして得られた音響モデル用音声データを使用して、特定話者の音響モデル２７ｄに対する単語認識率を実験により調べたところ、図７に示す結果が得られた。図７は、横軸に方向を，縦軸に単語認識率を示すグラフであり、符号Ｐは本人（話者Ａ）の音声，符号Ｑは他者（話者Ｂ，Ｃ）の音声の場合を示す。話者Ａの音響モデルでは、話者Ａがロボット１０の正面に位置している場合（図７（Ａ））には、正面（０度）にて８０％以上の単語認識率となり、また話者Ａが右方６０度または左方－６０度に位置する場合、それぞれ図７（Ｂ）又は図７（Ｃ）に示すように、話者よりも方向の違いによる認識率の低下が少なく、特に話者も方向もあっている場合には、８０％以上の単語認識率となることが分かった。

【００７６】

この結果を考慮して、音声認識の際に、音源方向が既知であることを利用して、セレクタ２７ｅは、以下の式（５）により与えられるコスト関数Ｖ（ｐｅ）を統合のために使用する。

【数５】

$$V(p_e)=\left(\sum_d r(p_e,d)\cdot v(p_e,d)+\sum_d r(p,d_e)\cdot v(p,d_e)-r(p_e,d_e)\right)\cdot P_v(p_e)$$

$$v(p,d)=\begin{cases}1 & if \quad \mathrm{Re}s(p,d)=\mathrm{Re}s(p_e,d_e)\\ 0 & if \quad \mathrm{Re}s(p,d)\neq\mathrm{Re}s(p_e,d_e)\end{cases}$$

【００７７】

ここで、ｒ（ｐ，ｄ），Ｒｅｓ（ｐ，ｄ）をそれぞれ話者ｐと方向ｄの音響モデルを使用した場合の単語認識率と入力音声に対する認識結果と定義し、d_e をリアルタイムトラッキングによる音源方向とし、さらにp_e を評価対象の話者とする。

【００７８】

上記P_v（p_e，d_e）は顔認識モジュールで生成される確率であり、顔認識ができない場合には常に１．０とする。そして、セレクタ２７ｅは最も大きいコスト関数Ｖ（p_e）を有する話者p_eと認識結果Ｒｅｓ（ｐ，ｄ）を出力する。

その際、セレクタ２７ｅは、顔モジュール３０からの顔認識による顔イベント３９を参照することにより、話者を特定することができるので、音声認識のロバスト性を向上させることができる。

【００７９】

なお、コスト関数Ｖ（p_e）の最大値が１．０以下または二番目に大きい値と近い場合には、音声認識が失敗または一つの候補に絞りきれなかったことにより音声認識ができないと判断して、その旨を後述する対話部２８に出力する。上記対話部２８は、対話制御部２８ａと音声合成部２８ｂとスピーカ２８ｃとから構成されている。上記対話制御部２８ａは、後述するアソシエーションモジュール６０により制御されることにより、音声認識部２７からの音声認識結果、即ち話者p_eと認識結果Ｒｅｓ（ｐ，ｄ）とに基づいて、対象とする話者に対する音声データを生成し、音声合成部２８ｂに出力する。上記音声合成部２８ｂは、対話制御部２８ａからの音声データに基づいてスピーカ２８ｃを駆動して、音声データに対応する音声を発する。

【００８０】

これにより、対話部２８は音声認識部２７からの音声認識結果に基づいて、例えば話者Ａが好きな数字として「１」と言った場合に、ロボット１０が当該話者Ａに正対した状態で、当該話者Ａに対して「Ａさんは「１」と言いました」というように音声を発することになる。

【００８１】

なお、対話部２８は、音声認識部２７から音声認識ができなかった旨が出力された場合には、ロボット１０が当該話者Ａに正対した状態で、当該話者Ａに対して、「あなたは「２ですか？４ですか？」と質問して、再度話者Ａの回答について音声認識を行なうようになっている。この場合、話者Ａに対してロボット１０が正対していることから、音声認識の精度がより一層向上することになる。

【００８２】

このようにして、聴覚モジュール２０は、マイク１６からの音響信号に基づいて、ピッチ抽出，音源の分離及び定位から少なくとも一人の話者を特定（話者同定）してその聴覚イベントを抽出し、ネットワークを介してアソシエーションモ

ジュール５０に対して送信すると共に、各話者の音声認識を行なって対話部２８により音声認識結果を話者に対して音声により確認するようになっている。

【００８３】

ここで、実際には、音源方向θ_Sが時間ｔの関数であることから、特定音源を抽出し続けるためには時間方向の連続性を考慮する必要があるが、上述したように、リアルタイムトラッキングからのストリーム方向θ_Sにより、音源方向を得るようにしている。

【００８４】

これによって、リアルタイムトラッキングにて、すべてのイベントをストリームという時間的流れを考慮した表現で表わしているので、同時に複数の音源が存在したり、音源やロボット自身が移動する場合でも、一つのストリームに注目することによって特定音源からの方向情報を連続的に得ることができる。さらに、ストリームは視聴覚のイベントを統合するためにも使用しているので、顔イベントを参照して聴覚イベントにより音源定位を行なうことにより、音源定位の精度が向上することになる。

【００８５】

上記顔モジュール３０は、デバイス層としてのカメラ１５と、プロセス層としての顔発見部３１，顔識別部３２，顔定位部３３と、特徴層（データ）としての顔ＩＤ３４，顔方向３５と、イベント層としての顔イベント生成部３６と、から構成されている。

【００８６】

これにより、顔モジュール３０は、カメラ１５からの画像信号に基づいて、顔発見部３１により例えば肌色抽出により各話者の顔を検出し、顔識別部３２にて前もって登録されている顔データベース３８により検索して、一致した顔があった場合、その顔ＩＤ３４を決定して当該顔を識別すると共に、顔定位部３３により当該顔方向３５を決定（定位）する。

【００８７】

ここで、顔モジュール３０は、顔発見部３１が画像信号から複数の顔を見つけた場合、各顔について上記処理、即ち識別及び定位そして追跡を行なう。その際

、顔発見部３１により検出された顔の大きさ，方向及び明るさがしばしば変化するので、顔発見部３１は顔領域検出を行なって、肌色抽出と相関演算に基づくパターンマッチングの組合せによって２００m秒以内に複数の顔を正確に検出できるようになっている。

【００８８】

顔定位部３３は、二次元の画像平面における顔位置を三次元空間に変換し、三次元空間における顔位置を、方位角θ，高さφ及び距離ｒのセットとして得る。そして、顔モジュール３０は、各顔毎に、顔ＩＤ（名前）３４及び顔方向３５から、顔イベント生成部３６により顔イベント３９を生成して、ネットワークを介してアソシエーションモジュール５０に対して送信するようになっている。

【００８９】

上記顔ステレオモジュール３７は、デバイス層としてのカメラ１５と、プロセス層としての視差画像生成部３７ａ，目標抽出部３７ｂと、特徴層（データ）としての目標方向３７ｃと、イベント層としてのステレオイベント生成部３７ｄとから構成されている。これにより、ステレオモジュール３７は、カメラ１５からの画像信号に基づいて視差画像生成部３７ａによって双方のカメラ１５の画像信号から視差画像を生成する。次いで、目標抽出部３７ｂが、視差画像を領域分割し、その結果、縦に長い物体が発見されれば、目標抽出部３７ｂはそれを人物候補として抽出し、その目標方向３７ｃを決定（定位）する。ステレオイベント生成部３７ｄは、目標方向３７ｃに基づいてステレオイベント３９ａを生成し、ネットワークを介してアソシエーションモジュール５０に対して送信するようになっている。

【００９０】

上記モータ制御モジュール４０は、デバイス層としてのモータ４１及びポテンショメータ４２と、プロセス層としてのＰＷＭ制御回路４３，ＡＤ変換回路４４及びモータ制御部４５と、データである特徴層としてのロボット方向４６と、イベント層としてのモータイベント生成部４７とから構成されている。これにより、モータ制御モジュール４０においては、モータ制御部４５がアテンション制御モジュール５７（後述）からの指令に基づいてＰＷＭ制御回路４３を介してモー

タ４１を駆動制御する。また、モータ４１の回転位置をポテンショメータ４２により検出する。この検出結果は、ＡＤ変換回路４４を介してモータ制御部４５に送られる。そして、モータ制御部４５は、ＡＤ変換回路４４から受け取った信号からロボット方向４６を抽出する。モータイベント生成部４７は、ロボット方向４６に基づいて、モータ方向情報から成るモータイベント４８を生成して、ネットワークを介してアソシエーションモジュール５０に対して送信するようになっている。

【００９１】

上記アソシエーションモジュール５０は、上述した聴覚モジュール２０，顔モジュール３０，ステレオモジュール３７，モータ制御モジュール４０に対して、階層的に上位に位置付けられており、各モジュール２０，３０，３７，４０のイベント層の上位であるストリーム層を構成している。具体的には、アソシエーションモジュール５０は、聴覚モジュール２０，顔モジュール３０，ステレオモジュール３７及びモータ制御モジュール４０からの非同期イベント５１、即ち聴覚イベント２９，顔イベント３９，ステレオイベント３９ａ及びモータイベント４８を同期させて聴覚ストリーム５３，顔ストリーム５４，ステレオ視覚ストリーム５５を生成する絶対座標変換部５２と、各ストリーム５３，５４，５５を関連付けてアソシエーションストリーム５９を生成し、あるいはこれらストリーム５３，５４，５５の関連付けを解除する関連付け部５６と、さらにアテンション制御モジュール５７と、ビューア５８とを備えている。

【００９２】

上記絶対座標変換部５２は、聴覚モジュール２０からの聴覚イベント２９，顔モジュール３０からの顔イベント３９，ステレオモジュール３７からのステレオイベント３９ａに、モータ制御モジュール４０からのモータイベント４８を同期させると共に、聴覚イベント２９，顔イベント３９及びステレオイベント３９ａに関して、同期させたモータイベントによって、その座標系を絶対座標系に変換することにより、聴覚ストリーム５３，顔ストリーム５４及びステレオ視覚ストリーム５５を生成する。その際、上記絶対座標変換部５２は、同一話者の聴覚ストリーム，顔ストリーム及びステレオ視覚ストリームに接続することによって、

聴覚ストリーム５３，顔ストリーム５４及びステレオ視覚ストリーム５５を生成する。

【００９３】

また、関連付け部５６は、聴覚ストリーム５３，顔ストリーム５４，ステレオ視覚ストリーム５５に基づいて、これらのストリーム５３，５４，５５の時間的つながりを考慮してストリームを関連付け、あるいは関連付けを解除して、アソシエーションストリーム５９を生成すると共に、逆にアソシエーションストリーム５９を構成する聴覚ストリーム５３，顔ストリーム５４及びステレオ視覚ストリーム５５の結び付きが弱くなれば、関係付けを解除するようになっている。これにより、目標となる話者が移動している場合であっても、当該話者の移動を予測して、その移動範囲となる角度範囲内であれば、上述したストリーム５３，５４，５５の生成を行なうことによって、当該話者の移動を予測して追跡できることになる。

【００９４】

また、アテンション制御モジュール５７は、モータ制御モジュール４０の駆動モータ制御のプランニングのためのアテンション制御を行なうものであり、その際アソシエーションストリーム５９，聴覚ストリーム５３，顔ストリーム５４そしてステレオ視覚ストリーム５５の順に優先的に参照して、アテンション制御を行なう。そして、アテンション制御モジュール５７は、聴覚ストリーム５３，顔ストリーム５４及びステレオ視覚ストリーム５５の状態とアソシエーションストリーム５９の存否に基づいて、ロボット１０の動作プランニングを行ない、駆動モータ４１の動作の必要があれば、モータ制御モジュール４０に対して動作指令としてのモータイベントをネットワークを介して送信する。ここで、アテンション制御モジュール５７におけるアテンション制御は、連続性とトリガに基づいており、連続性により同じ状態を保持しようとし、トリガにより最も興味のある対象を追跡しようとして、アテンションを向けるべきストリームを選択して、トラッキングを行なう。

【００９５】

このようにして、アテンション制御モジュール５７は、アテンション制御を行

なって、モータ制御モジュール４０の駆動モータ４１の制御のプランニングを行ない、このプランニングに基づいて、モータコマンド６４ａを生成し、ネットワーク７０を介してモータ制御モジュール４０に伝送する。これにより、モータ制御モジュール４０では、このモータコマンド６４ａに基づいて、モータ制御部４５がＰＷＭ制御を行なって、駆動モータ４１を回転駆動させて、ロボット１０を所定方向に向けるようになっている。

【００９６】

ビューア５８は、このようにして生成された各ストリーム５３，５４，５５，５７をサーバの画面上に表示するものであり、具体的にはレーダチャート５８ａ及びストリームチャート５８ｂにより表示する。レーダチャート５８ａは、その瞬間におけるストリームの状態、より詳細にはカメラの視野角と音源方向を示し、ストリームチャート５８ｂは、アソシエーションストリーム（太線図示）と聴覚ストリーム，顔ストリーム及びステレオ視覚ストリーム（細線図示）を示している。

【００９７】

本発明実施形態による人型ロボット１０は以上のように構成されており、以下のように動作する。

まず、ロボット１０の前方１ｍの距離で、斜め左（θ＝＋６０度），正面（θ＝０度）そして斜め右（θ＝－６０度）の方向に、それぞれ話者が並んでおり、ロボット１０が対話部２８により、三人の話者に質問して、各話者が同時に質問に対する回答を行なう。

【００９８】

これにより、ロボット１０はマイク１６が当該話者の音声を拾って、聴覚モジュール２０が音源方向を伴う聴覚イベント２９を生成して、ネットワークを介してアソシエーションモジュール５０に伝送する。これにより、アソシエーションモジュール５０は、この聴覚イベント２９に基づいて、聴覚ストリーム５３を生成する。

【００９９】

また、顔モジュール３０は、カメラ１５による話者の顔の画像を取り込んで、

顔イベント３９を生成して、当該話者の顔を顔データベース３８により検索し、顔識別を行なうと共に、その結果である顔ＩＤ２４及び画像をネットワーク７０を介してアソシエーションモジュール５０に伝送する。なお、当該話者の顔が顔データベース３８に登録されていない場合には、顔モジュール３０は、その旨をネットワークを介してアソシエーションモジュール５０に伝送する。

【０１００】

従って、アソシエーションモジュール５０は、これらの聴覚イベント２９，顔イベント３９，ステレオイベント３９ａに基づいて、アソシエーションストリーム５９を生成する。

【０１０１】

ここで、聴覚モジュール２０は、アクティブ方向通過型フィルタ２３ａにより、聴覚エピポーラ幾何によるＩＰＤを利用して、各音源（話者Ｘ，Ｙ，Ｚ）の定位及び分離を行なって、分離音（音響信号）を取り出す。そして、聴覚モジュール２０は、その音声認識部２７により音声認識エンジン２７ｃを使用して、各話者Ｘ，Ｙ，Ｚの音声を認識してその結果を対話部２８に出力する。これにより、対話部２８は、音声認識部２７により音声認識された前記回答を、それぞれの話者に対してロボット１０が正対した状態で発話する。なお、音声認識部２７が正しく音声認識できなかった場合には、ロボット１０が当該話者に正対した状態で再度質問を繰り返し、その回答に基づいて再度音声認識を行なう。

【０１０２】

このようにして、本発明実施形態による人型ロボット１０によれば、聴覚モジュール２０により音源定位・音源分離された分離音（音響信号）に基づいて、音声認識部２７が、各話者及び方向に対応する音響モデルを使用して音声認識を行なうことにより同時に発話する複数の話者の音声を音声認識することができる。

【０１０３】

以下に、音声認識部２７の動作を実験により評価する。

これらの実験においては、図８に示すように、ロボット１０の前方１ｍの距離で、斜め左（θ＝＋６０度），正面（θ＝０度）そして斜め右（θ＝－６０度）の方向に、それぞれ話者Ｘ，Ｙ，Ｚが並んでいる。なお、実験では、話者として

人間の代わりにそれぞれスピーカを置くと共に、その前面に話者の写真を配置している。ここで、スピーカは、音響モデルを作成したときと同じスピーカを使用しており、スピーカから発せられた音声を写真の話者の音声とみなしている。

【０１０４】

そして、以下のシナリオに基づいて音声認識の実験を行なう。

１．ロボット１０が三人の話者Ｘ，Ｙ，Ｚに質問する。

２．三人の話者Ｘ，Ｙ，Ｚが同時に質問に対する回答を行なう。

３．ロボット１０が三人の話者Ｘ，Ｙ，Ｚの混合音声に基づいて、音源定位・音源分離を行ない、さらに各分離音について音声認識を行なう。

４．ロボット１０が、順次に各話者Ｘ，Ｙ，Ｚに正対した状態で当該話者の回答を答える。

５．ロボット１０は、音声認識が正しくできなかったと判断したとき、当該話者に正対して再度質問を繰り返し、その回答に基づいて再度音声認識を行なう。

【０１０５】

上記シナリオによる実験結果の第一の例を図９に示す。

１．ロボット１０が「好きな数字は何ですか？」と質問する。（図９（ａ）参照）

２．各話者Ｘ，Ｙ，Ｚとしてのスピーカから、同時に１から１０までの数字のうちから、任意の数字を読み上げた音声を流す。例えば図９（ｂ）に示すように、話者Ｘは「２」，話者Ｙは「１」そして話者Ｚは「３」と言う。

３．ロボット１０は、聴覚モジュール２０にて、そのマイク１６で集音した音響信号に基づいて、アクティブ方向通過型フィルタ２３ａにより音源定位・音源分離を行なって、分離音を抽出する。そして、各話者Ｘ，Ｙ，Ｚに対応する分離音に基づいて、各話者別に音声認識部２７が９つの音響モデルを使用して、同時に音声認識プロセスを実行し、その音声認識を行なう。

４．その際、音声認識部２７のセレクタ２７ｅが、正面が話者Ｙであると仮定して音声認識の評価を行ない（図９（ｃ））、続いて正面が話者Ｘであると仮定して音声認識の評価を行ない（図９（ｄ））、最後に正面が話者Ｚであると仮定して音声認識の評価を行なう（図９（ｅ））。

５．そして、セレクタ２７ｅが、音声認識結果を統合して、図９（ｆ）に示すように、ロボット正面（θ＝０度）に関して、最も適合の良い話者名（Ｙ）と音声認識結果（「１」）を決定し対話部２８に出力する。これにより、図９（ｇ）に示すように、ロボット１０が話者Ｙに正対した状態にて、「Ｙさんは「１」です。」と答える。

６．続いて、斜め左（θ＝＋６０度）の方向に関して、上記と同様の処理を行って、図９（ｈ）に示すように、ロボット１０が話者Ｘに正対した状態にて、「Ｘさんは「２」です。」と答える。更に、斜め右（θ＝－６０度）の方向に対しても同様の処理を行って、図９（ｉ）に示すように、ロボット１０が話者Ｚに正対した状態にて、「Ｚさんは「３」です。」と答える。

【０１０６】

この場合、ロボット１０は、各話者Ｘ，Ｙ，Ｚの回答をすべて正しく音声認識することができた。従って、同時発話の場合であっても、ロボット１０のマイク１６を使用したロボット視聴覚システム１７における音源定位・音源分離・音声認識の有効性が示された。

【０１０７】

なお、図９（ｊ）に示すように、ロボット１０が各話者に正対せずに、「Ｙさんは「１」です。Ｘさんは「２」です。Ｚさんは「３」です。合計「６」です。」というように、各話者Ｘ，Ｙ，Ｚの答えた数字の合計も答えるようにしてもよい。

【０１０８】

図１０は、上述したシナリオによる実験結果の第二の例を示している。

１．図９に示した第一の例と同様にして、ロボット１０が「好きな数字は何ですか？」と質問し（図１０（ａ）参照）、各話者Ｘ，Ｙ，Ｚとしてのスピーカから、図１０（ｂ）に示すように、話者Ｘは「２」，話者Ｙは「１」そして話者Ｚは「３」という音声が流れる。

２．ロボット１０は、同様にして、聴覚モジュール２０にて、そのマイク１６で集音した音響信号に基づいて、アクティブ方向通過型フィルタ２３ａにより音源定位・音源分離を行なって分離音を抽出し、各話者Ｘ，Ｙ，Ｚに対応する分離

音に基づいて、各話者別に音声認識部２７が９つの音響モデルを使用して、同時に音声認識プロセスを実行し、その音声認識を行なう。その際、音声認識部２７のセレクタ２７ｅは、図１０（ｃ）に示すように、正面の話者Ｙについては、正しく音声認識の評価を行なうことができる。

３．これに対して、＋６０度に位置する話者Ｘについて、セレクタ２７ｅは、図１０（ｄ）に示すように、「２」であるか「４」であるか決定することができない。

４．従って、ロボット１０は、図１０（ｅ）に示すように、＋６０度に位置する話者Ｘに正対して、「２ですか？　４ですか？」と質問する。

５．これに対して、図１０（ｆ）に示すように、話者Ｘであるスピーカから「２」という回答が流れる。この場合、話者Ｘは、ロボット１０の正面に位置していることから、聴覚モジュール２０が話者Ｘの回答について正しく音源定位・音源分離し、音声認識部２７が正しく音声認識して、話者名Ｘと音声認識結果「２」を対話部２８に出力する。これにより、ロボット１０は、図１０（ｇ）に示すように、話者Ｘに対して「Ｘさんは「２」です。」と答える。

６．続いて、話者Ｚについても同様の処理を行なって、その音声認識結果を話者Ｚに対して答える。即ち、図１０（ｈ）に示すように、ロボット１０が話者Ｚに正対した状態にて、「Ｚさんは「３」です。」と答える。

【０１０９】

このようにして、ロボット１０は、再質問により、各話者Ｘ，Ｙ，Ｚの回答をすべて正しく音声認識することができた。従って、側方での聴覚中心窩の影響による分離精度の低下による音声認識の曖昧さを、ロボット１０が側方の話者に対して正対して再質問することにより解消して、音源分離精度を向上させ、音声認識精度を向上させることができることが示された。

【０１１０】

なお、図１０（ｉ）に示すように、ロボット１０が各話者の音声認識を正しく行なった後、「Ｙさんは「１」です。Ｘさんは「２」です。Ｚさんは「３」です。合計「６」です。」というように、各話者Ｘ，Ｙ，Ｚの答えた数字の合計も答えるようにしてもよい。

【０１１１】

図１１は、上述したシナリオによる実験結果の第三の例を示している。

１．この場合も図９に示した第一の例と同様にして、ロボット１０が「好きな数字は何ですか？」と質問し（図１０（ａ）参照）、各話者Ｘ，Ｙ，Ｚとしてのスピーカから、図１０（ｂ）に示すように、話者Ｘは「８」，話者Ｙは「７」そして話者Ｚは「９」という音声が流れる。

２．ロボット１０は、同様にして、聴覚モジュール２０にて、そのマイク１６で集音した音響信号に基づいて、リアルタイムトラッキング（Ｘ３′参照）によるストリーム方向θ、そして各話者の顔イベントを参照して、アクティブ方向通過型フィルタ２３ａにより音源定位・音源分離を行なって分離音を抽出し、各話者Ｘ，Ｙ，Ｚに対応する分離音に基づいて、各話者毎に音声認識部２７が９つの音響モデルを使用して、同時に音声認識プロセスを実行し、その音声認識を行なう。

その際、音声認識部２７のセレクタ２７ｅは、正面の話者Ｙについては、顔イベントに基づいて話者Ｙである確率が高いことから、各音響モデルによる音声認識結果の統合の際に、図１０（ｃ）に示すようにこれを考慮する。これにより、より正確な音声認識を行なうことができる。従って、ロボット１０は、図１１（ｄ）に示すように、話者Ｘに対して「Ｘさんは「７」です。」と答える。

３．これに対して、＋６０度に位置する話者Ｘについて、ロボット１０が向きを変えて正対すると、このときの正面の話者Ｘについて、顔イベントに基づいて話者Ｘである確率が高いので、同様にして、セレクタ２７ｅは、図１１（ｅ）に示すようにこれを考慮する。従って、ロボット１０は、図１１（ｆ）に示すように、話者Ｘに対して「Ｙさんは「８」です。」と答える。

４．続いて、セレクタ２７ｅは、図１１（ｇ）に示すように、話者Ｚについても同様の処理を行なって、その音声認識結果を話者Ｚに対して答える。即ち、図１１（ｈ）に示すように、ロボット１０が話者Ｚに正対した状態にて「Ｚさんは「９」です。」と答える。

【０１１２】

このようにして、ロボット１０は、各話者毎に正対して、その顔イベントを参

照しながら話者の顔認識に基づいて、各話者Ｘ，Ｙ，Ｚの回答をすべて正しく音声認識することができた。これにより、顔認識により話者が誰であるかを特定することができるので、より精度の高い音声認識を行なうことができることが示された。特に、特定の環境での利用を前提とするような場合、顔認識によってほぼ１００％に近い顔認識精度が得られると、顔認識情報を信頼性の高い情報として利用することができることになり、音声認識部２７の音声認識エンジン２７ｃで使用される音響モデル２７ｄの数を削減することができるので、より高速で且つ高精度の音声認識が可能になる。

【０１１３】

図１２は、上述したシナリオによる実験結果の第四の例を示している。

１．ロボット１０が「好きなフルーツは何ですか？」と質問し（図１２（ａ）参照）、各話者Ｘ，Ｙ，Ｚとしてのスピーカから、例えば図１２（ｂ）に示すように、話者Ｘは「梨」，話者Ｙは「スイカ」そして話者Ｚは「メロン」と言う。

２．ロボット１０は、聴覚モジュール２０にて、そのマイク１６で集音した音響信号に基づいて、アクティブ方向通過型フィルタ２３ａにより音源定位・音源分離を行なって分離音を抽出する。そして、各話者Ｘ，Ｙ，Ｚに対応する分離音に基づいて、各話者毎に音声認識部２７が９つの音響モデルを使用して、同時に音声認識プロセスを実行し、その音声認識を行なう。

３．その際、音声認識部２７のセレクタ２７ｅが、正面が話者Ｙであると仮定して音声認識の評価を行ない（図１２（ｃ））、続いて正面が話者Ｘであると仮定して音声認識の評価を行ない（図１２（ｄ））、最後に正面が話者Ｚであると仮定して音声認識の評価を行なう（図１２（ｅ））。

４．そして、セレクタ２７ｅが、音声認識結果を統合して、図１２（ｆ）に示すように、ロボット正面（θ＝０度）方向に関して最も適合の良い話者名（Ｙ）と音声認識結果（「スイカ」）を決定し対話部２８に出力する。これにより、図９（ｇ）に示すように、ロボット１０が話者Ｙに正対した状態にて、「Ｙさんは「スイカ」です。」と答える。

５．続いて、各話者Ｘ，Ｚについても同様の処理を行なって、その音声認識結果を各話者Ｘ，Ｚに対して答える。即ち、図１２（ｈ）に示すように、ロボット

１０が話者Ｘに正対した状態にて、「Ｘさんは「梨」です。」と答え、さらに図１２（ｉ）に示すように、ロボット１０が話者Ｚに正対した状態にて「Ｚさんは「メロン」です。」と答える。

【０１１４】

この場合、ロボット１０は、各話者Ｘ，Ｙ，Ｚの回答をすべて正しく音声認識することができた。従って、音声認識エンジン２７ｃに登録された単語は数字に限ることなく、前もって登録された単語であれば、音声認識可能であることが分かる。ここで、実験に使用した音声認識エンジン２７ｃでは、約１５０語の単語が登録されている。なお、単語の音節数が多くなると、音声認識率はやや低くなる。

【０１１５】

上述した実施形態においては、ロボット１０は、その上半身が４ＤＯＦ（自由度）を有するように構成されているが、これに限らず、任意の動作を行なうように構成されたロボットに本発明によるロボット視聴覚システムを組み込むことも可能である。

【０１１６】

また、上述した実施形態においては、本発明によるロボット視聴覚システムを人型ロボット１０に組み込んだ場合について説明したが、これに限らず、犬型等の各種動物型ロボットや、その他の形式のロボットに組み込むことも可能であることは明らかである。

【０１１７】

また、上記説明では、図４に示すようにロボット視聴覚システム１７がステレオモジュール３７を備える構成例を説明したが、本発明の実施形態に係るロボット視聴覚システムは、ステレオモジュール３７を備えずに構成することもできる。この場合、アソシエーションモジュール５０は、聴覚イベント２９，顔イベント３９及びモータイベント４８に基づいて、各話者の聴覚ストリーム５３及び顔ストリーム５４を生成し、さらにこれらの聴覚ストリーム５３及び顔ストリーム５４を関連付けてアソシエーションストリーム５９を生成するように構成され、アテンション制御モジュール５０においては、これらのストリームに基づいてア

テンション制御が行われるように構成される。

【０１１８】

さらに、上記説明においては、アクティブ方向通過型フィルタ２３ａは、方向毎に通過帯域幅（パスレンジ）を制御しており、処理する音の周波数によらず通過帯域幅を一定としていた。

ここで、通過帯域δを導出するために、１００Ｈｚ，２００Ｈｚ，５００Ｈｚ，１０００Ｈｚ，２０００Ｈｚ，１００Ｈｚの調波構造音（ハーモニクス）の５つの純音と１つのハーモニクスとを用いて、１音源に対する音源抽出率を調べる実験を行った。なお、音源をロボット正面である０度からロボットの左位置或いは右位置である９０度の範囲で１０度毎に位置を移動させた。図１３～図１５は音源を０度から９０度の範囲の各位置に設置した場合の音源抽出率を示すグラフであり、この実験結果が示すように、周波数に応じて通過帯域幅を制御することにより、特定の周波数の音の抽出率を向上させることができ、分離精度を向上できる。よって、音声認識率も向上する。従って、上記説明したロボット視聴覚システム１７においては、アクティブ方向通過型フィルタ２３ａのパスレンジが、周波数毎に制御可能に構成されるのが望ましい。

【０１１９】

【発明の効果】

以上述べたように、この発明によれば、従来の音声認識と比較して、複数の音響モデルを使用することによって、実時間・実環境での正確な音声認識を行なうことが可能である。また、各音響モデルによる音声認識結果をセレクタにより統合して、最も信頼性の高い音声認識結果を判断することにより、従来の音声認識に比べて、より一層正確な音声認識を行なうことができる。

【図面の簡単な説明】

【図１】

この発明によるロボット聴覚装置の第一の実施形態を組み込んだ人型ロボットの外観を示す正面図である。

【図２】

図１の人型ロボットの側面図である。

【図３】

図１の人型ロボットにおける頭部の構成を示す概略拡大図である。

【図４】

図１の人型ロボットにおけるロボット視聴覚システムの電気的構成例を示すブロック図である。

【図５】

図４に示すロボット視聴覚システムにおける聴覚モジュールの作用を示す図である。

【図６】

図４のロボット視聴覚システムにおける聴覚モジュールの音声認識部で使用される音声認識エンジンの構成例を示す概略斜視図である。

【図７】

図６の音声認識エンジンによる正面及び左右±６０度の方向の話者による音声の認識率を示すグラフであり、（Ａ）は正面の話者、（Ｂ）は斜め左＋６０度の話者そして（Ｃ）は斜め右－６０度の話者の場合を示している。

【図８】

図４に示すロボット視聴覚システムにおける音声認識実験を示す概略斜視図である。

【図９】

図４のロボット視聴覚システムの音声認識実験の第一の例の結果を順次に示す図である。

【図１０】

図４のロボット視聴覚システムの音声認識実験の第二の例の結果を順次に示す図である。

【図１１】

図４のロボット視聴覚システムの音声認識実験の第三の例の結果を順次に示す図である。

【図１２】

図４のロボット視聴覚システムの音声認識実験の第四の例の結果を順次に示す

図である。

【図１３】

本発明の実施形態に係るアクティブ方向通過型フィルタの通過帯域幅を制御した場合の抽出率を示す図であり、（ａ）は０度、（ｂ）は１０度、（ｃ）は２０度、（ｄ）は３０度の方向に音源がある場合である。

【図１４】

本発明の実施形態に係るアクティブ方向通過型フィルタの通過帯域幅を制御した場合の抽出率を示す図であり、（ａ）は４０度、（ｂ）は５０度、（ｃ）は６０度の方向に音源がある場合である。

【図１５】

本発明の実施形態に係るアクティブ方向通過型フィルタの通過帯域幅を制御した場合の抽出率を示す図であり、（ａ）は７０度、（ｂ）は８０度、（ｃ）は９０度の方向に音源がある場合である。

【符号の説明】

１０　　人型ロボット
１１　　ベース
１２　　胴体部
１３　　頭部
１４　　外装
１５　　カメラ（ロボット視覚）
１６，１６ａ，１６ｂ　　マイク（ロボット聴覚）
１７　　ロボット視聴覚システム
２０　　聴覚モジュール
２１　　ピーク抽出部
２２　　音源定位部
２３　　音源分離部
２３ａ　アクティブ方向通過型フィルタ
２６　　聴覚イベント生成部
２７　　音声認識部

２７ａ　自声抑制部

２７ｂ　自動認識部

２７ｃ　音声認識エンジン

２７ｄ　音響モデル

２７ｅ　セレクタ

２８　　対話部

３０　　顔モジュール

３７　　ステレオモジュール

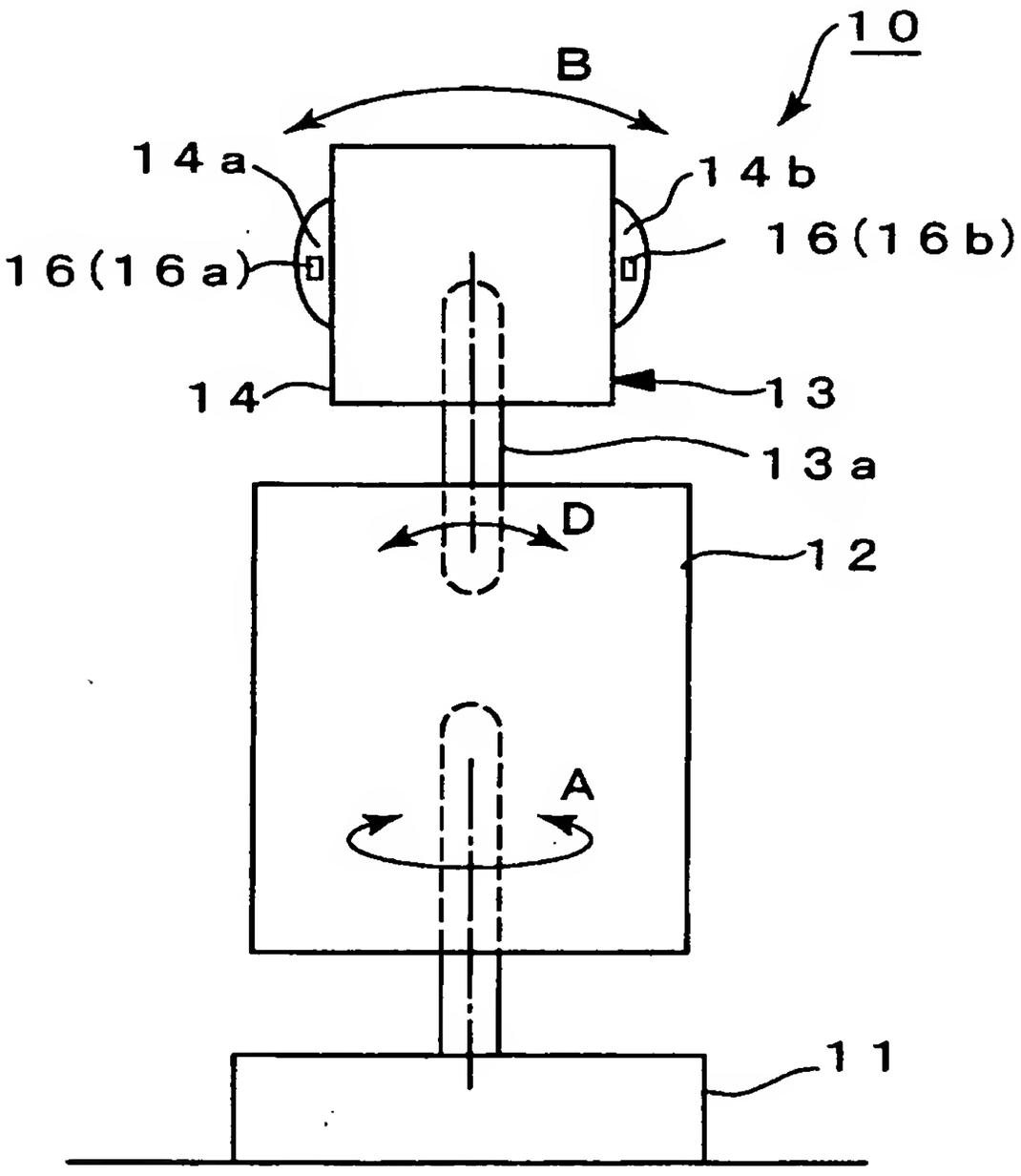
４０　　モータ制御モジュール

５０　　アソシエーションモジュール

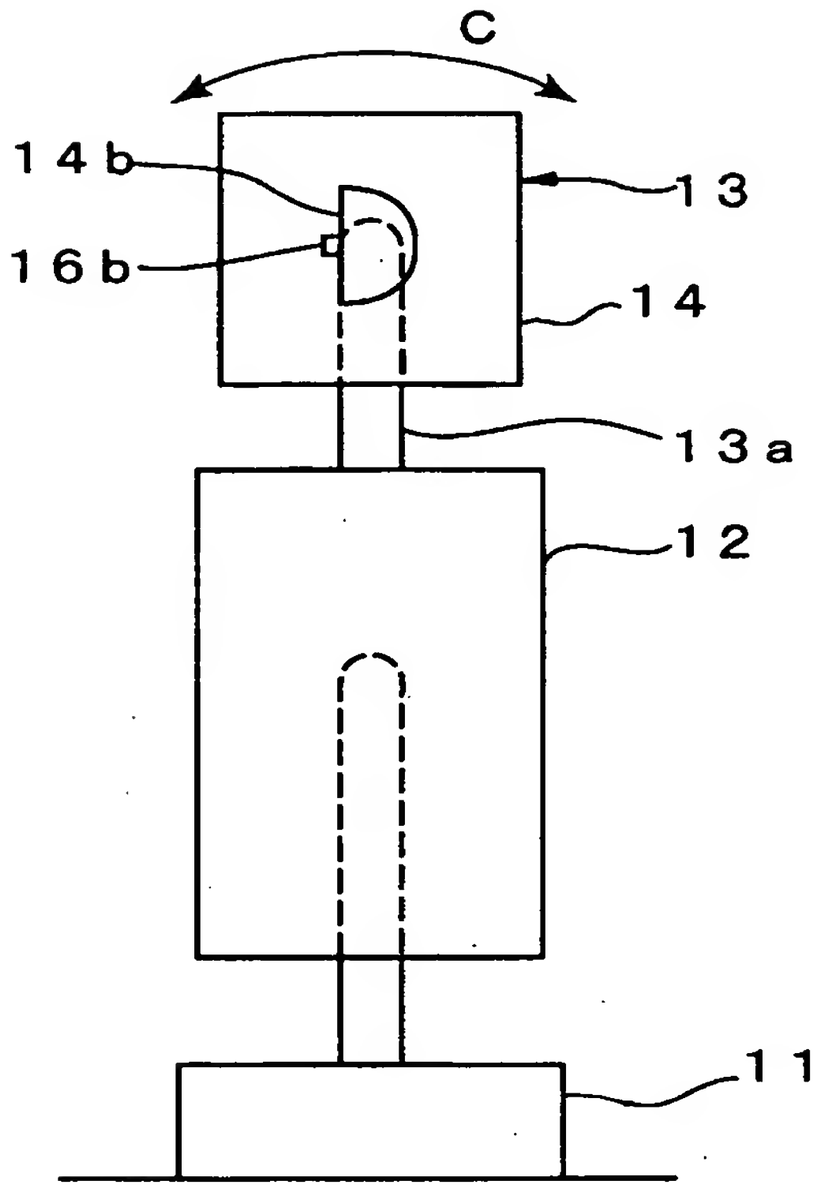
５７　　アテンション制御モジュール

【書類名】　図面

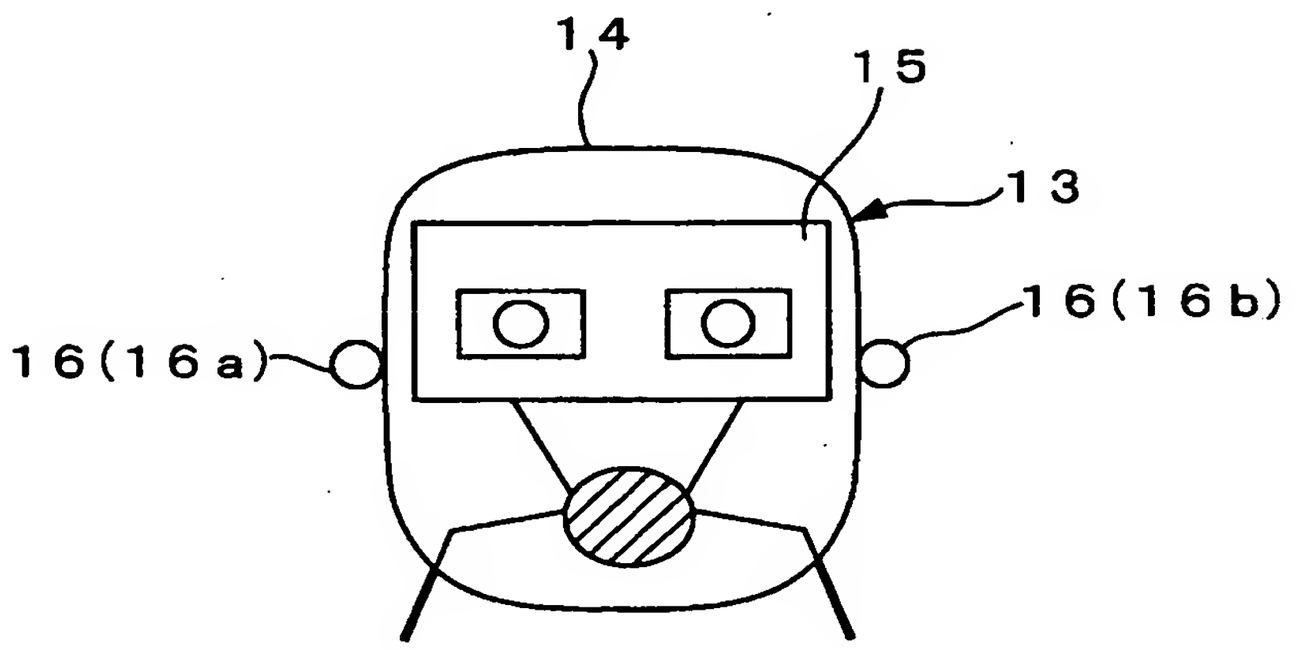
【図１】



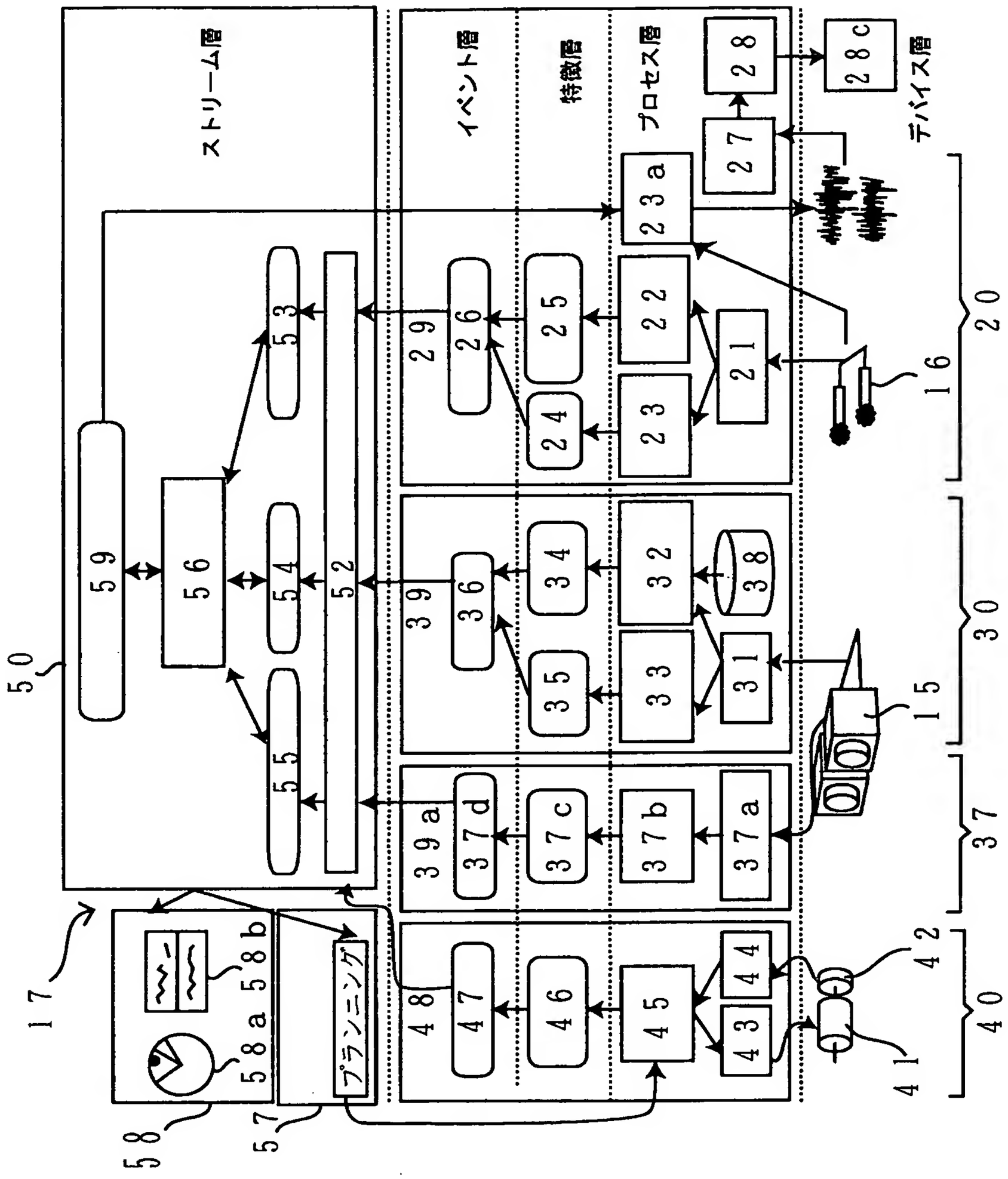
【図２】



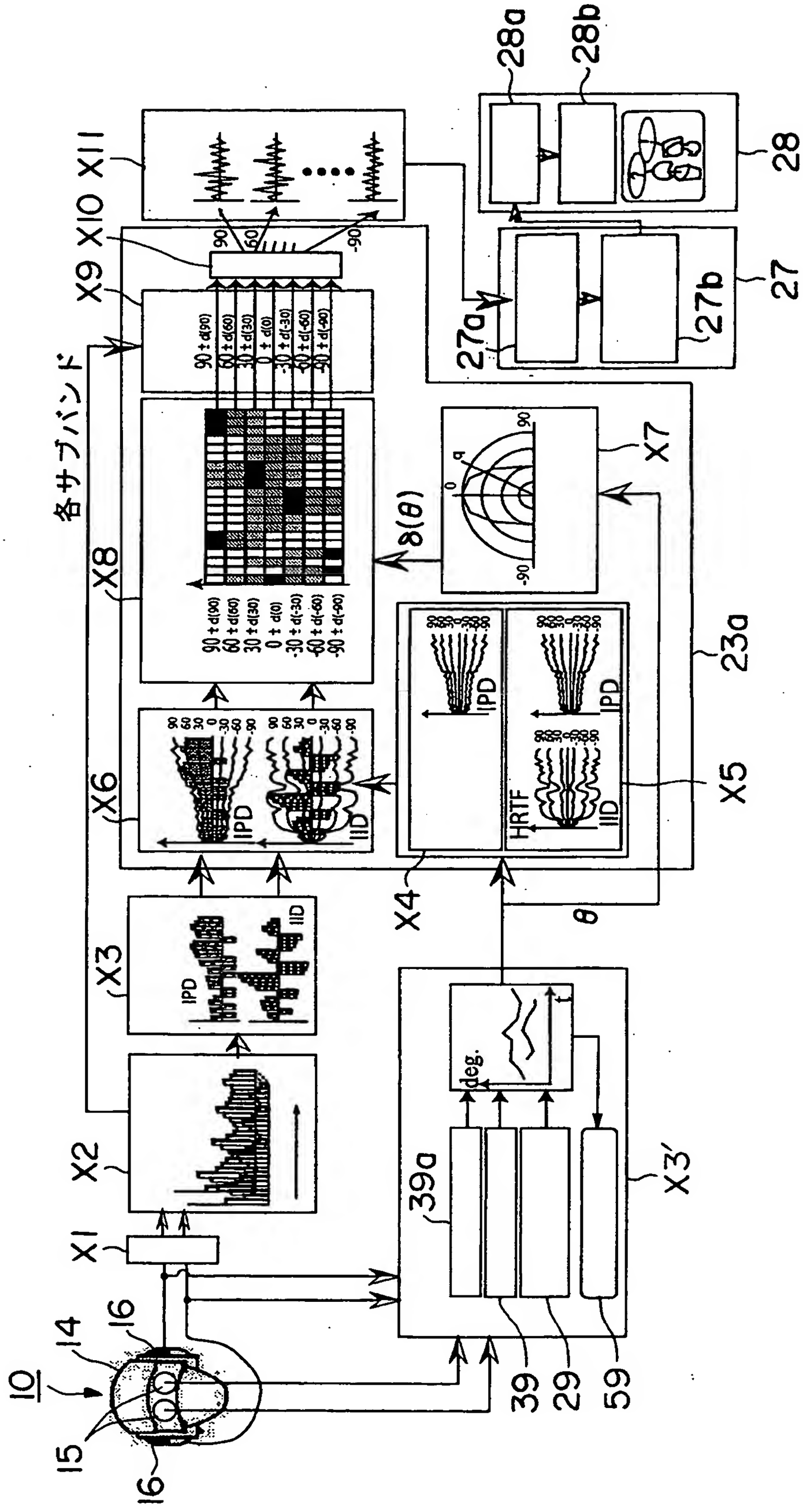
【図３】



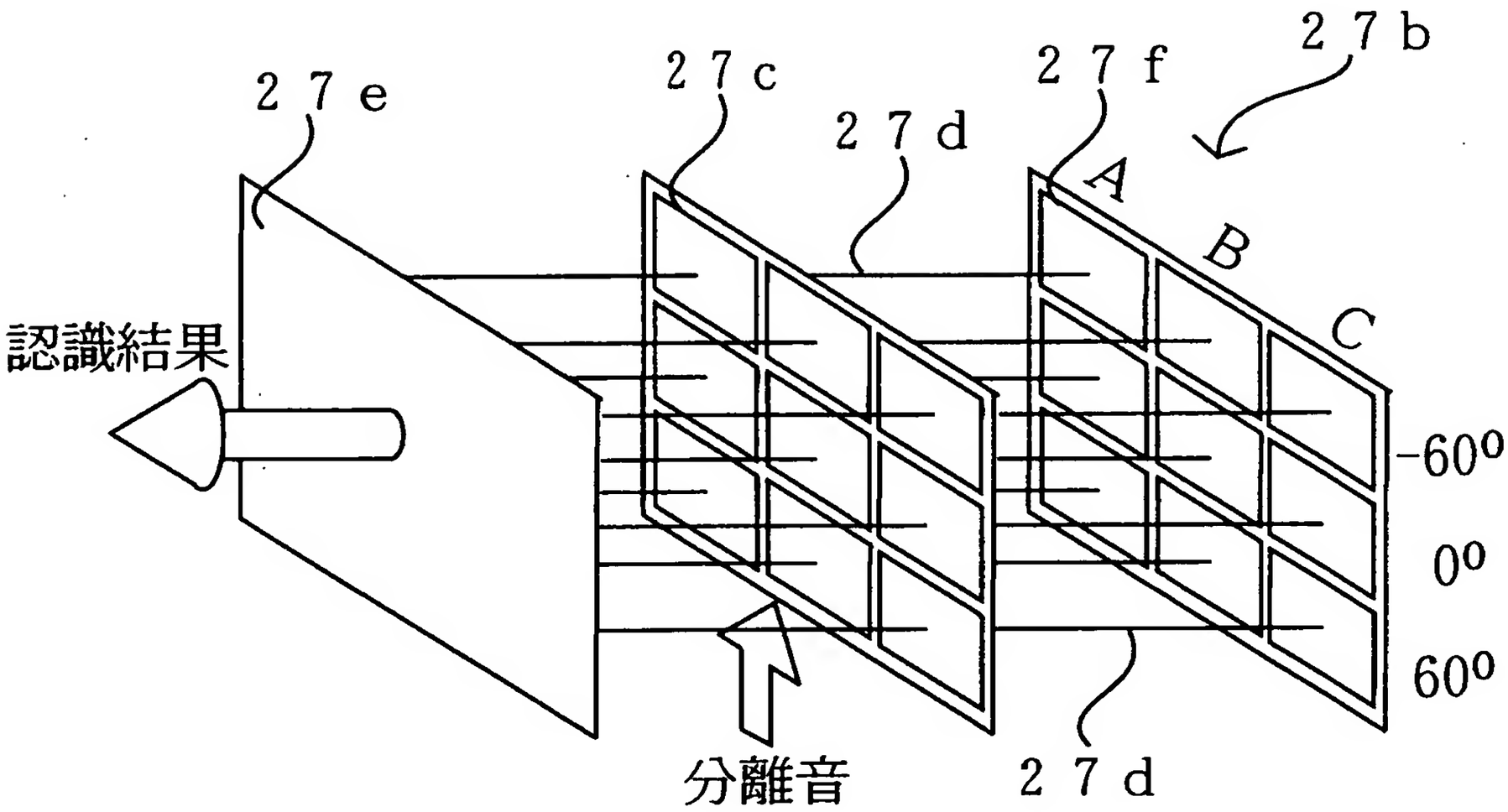
【図４】


ストリーム層
イベント層
特徴層
プロセス層
デバイス層
50
52
53
54
55
56
59
17
58a
58b
8
57
プランニング
29
26
25
24
23a
23
22
21
27
28
28c
16
20
39
36
34
35
32
33
31
38
15
30
39a
37d
37c
37b
37a
37
48
47
46
45
44
43
42
41
40

【図５】

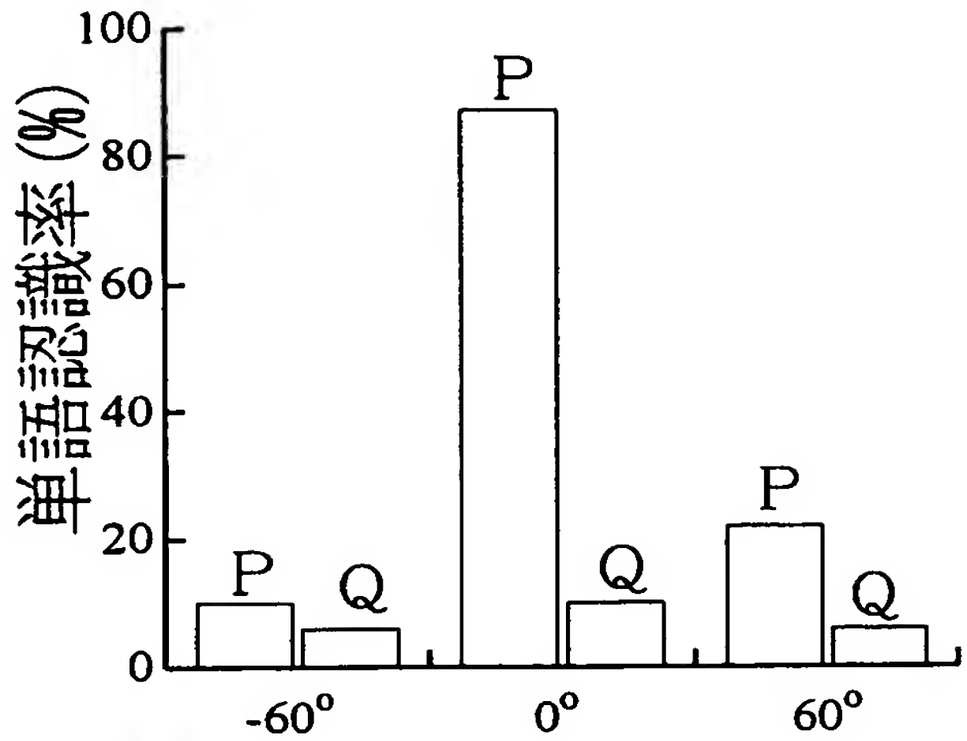

28a
28b
28
X9 X10 X11
90
60
-90
27a
27b
27
各サブバンド
90 ± d(90)
60 ± d(60)
30 ± d(30)
± d(0)
-30 ± d(-30)
-60 ± d(-60)
-90 ± d(-90)
X8
X7
δ(θ)
90
0
-90
X6
IPD
IID
X4
HRTF
X5
23a
X3
θ
deg.
X2
39a
X3'
X1
16
14
10
15
39
29
59

【図６】

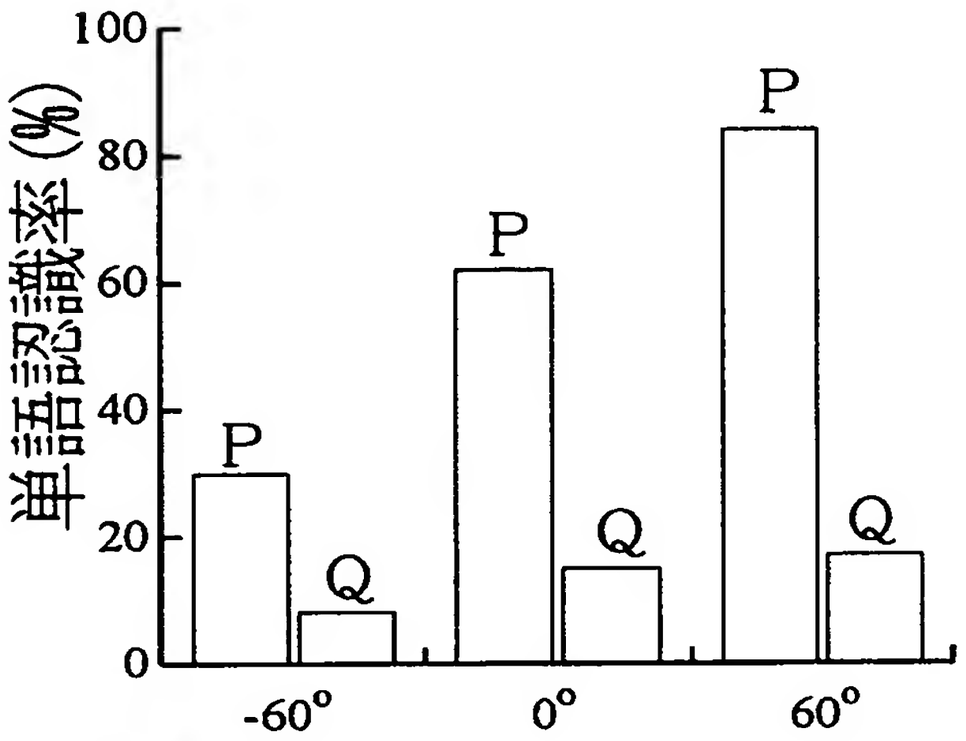


【図７】

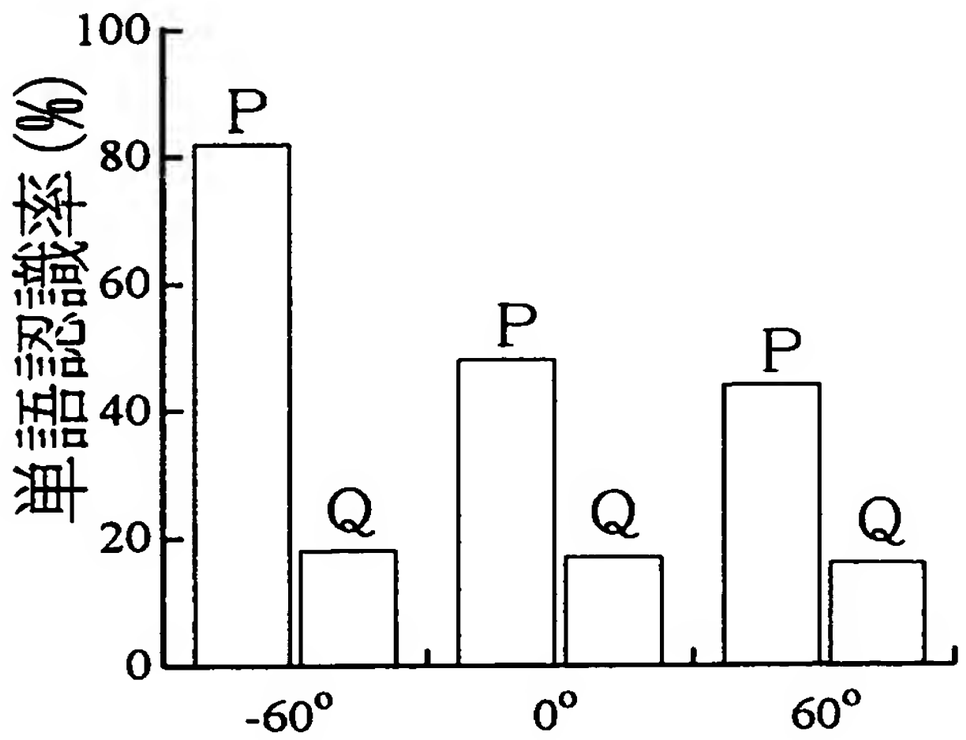
（A）


単語認識率（%）
100
80
60
40
20
0
P
Q
P
Q
P
Q
-60°
0°
60°

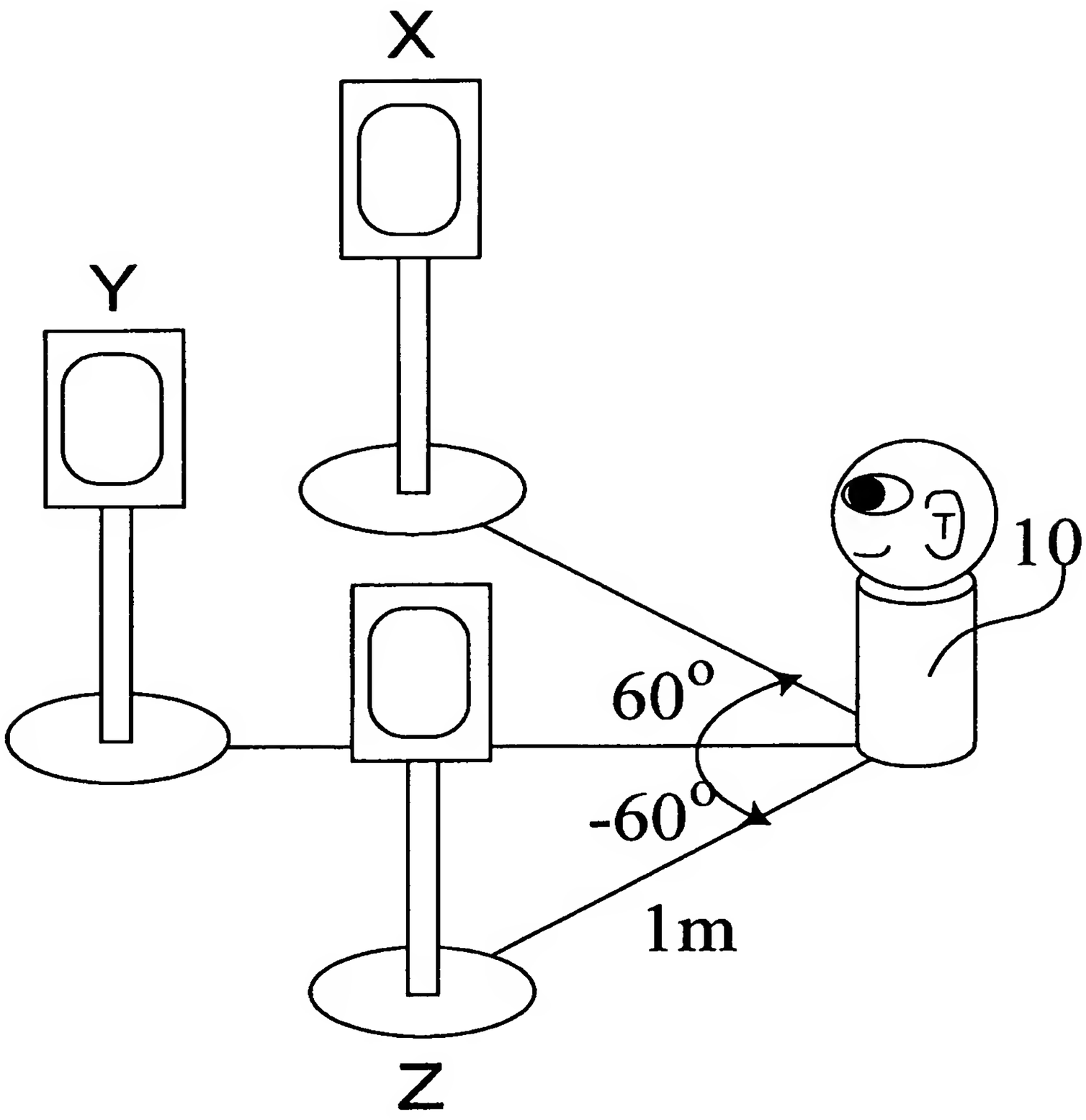
（B）


単語認識率（%）
100
80
60
40
20
0
P
Q
P
Q
P
Q
-60°
0°
60°

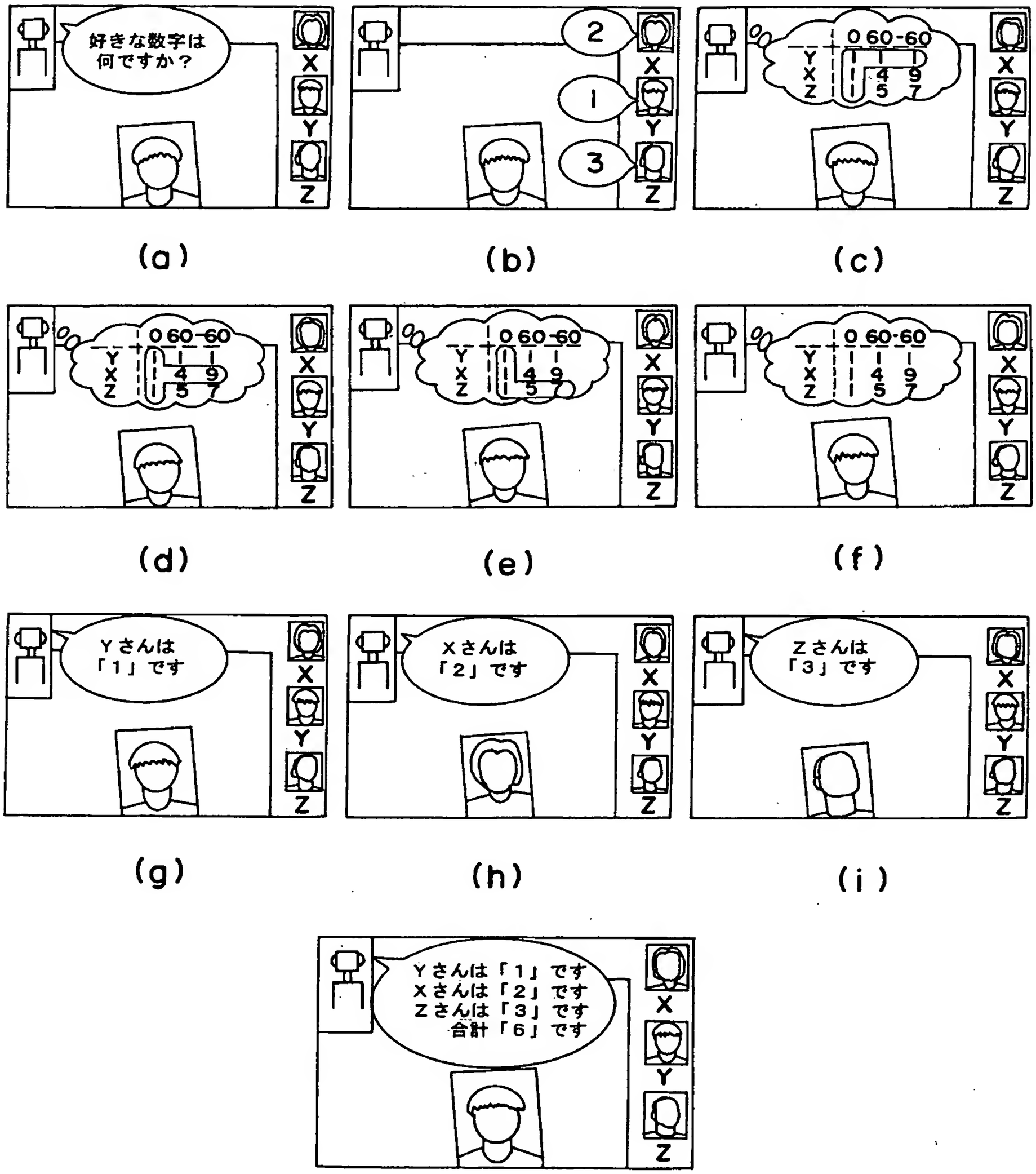
（C）


単語認識率（%）
100
80
60
40
20
0
P
Q
P
Q
P
Q
-60°
0°
60°

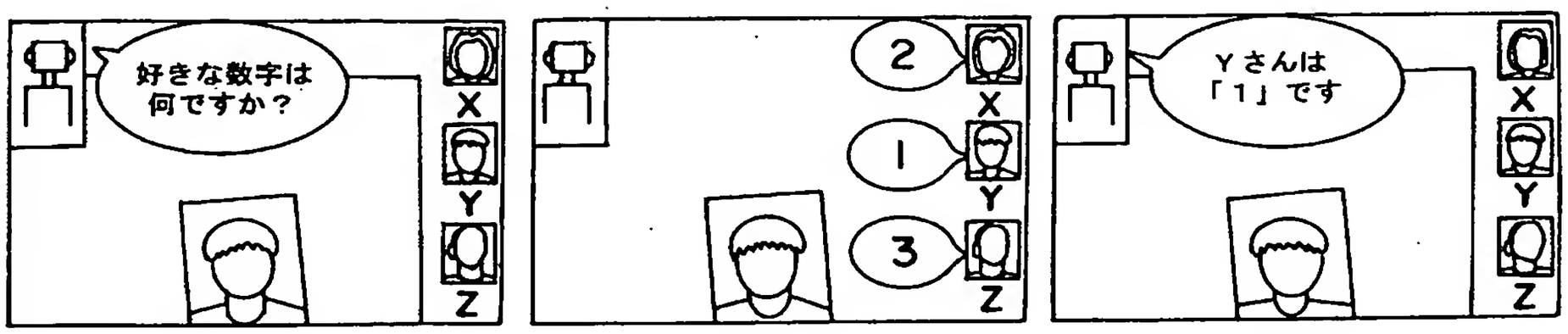
【図８】



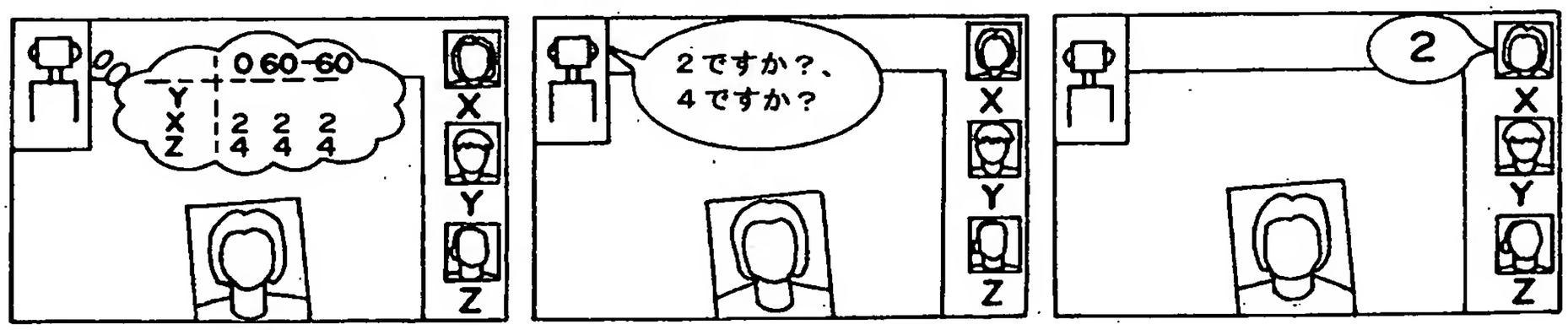
【図９】


好きな数字は
何ですか？
X
Y
Z
(a)
2
1
3
(b)
0 60-60
Y
X
Z
4 9
5 7
(c)
(d)
(e)
(f)
Yさんは「1」です
Xさんは「2」です
Zさんは「3」です
(g)
(h)
(i)
Yさんは「1」です
Xさんは「2」です
Zさんは「3」です
合計「6」です
(j)

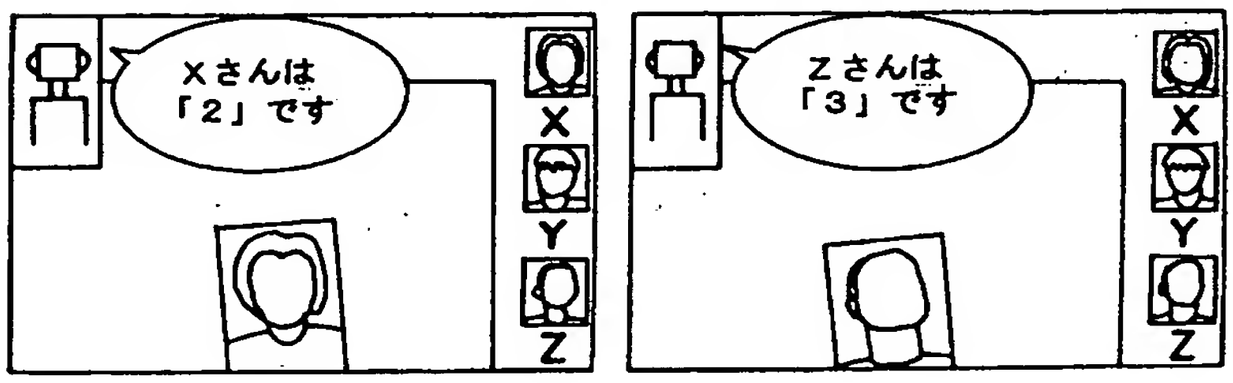
【図１０】



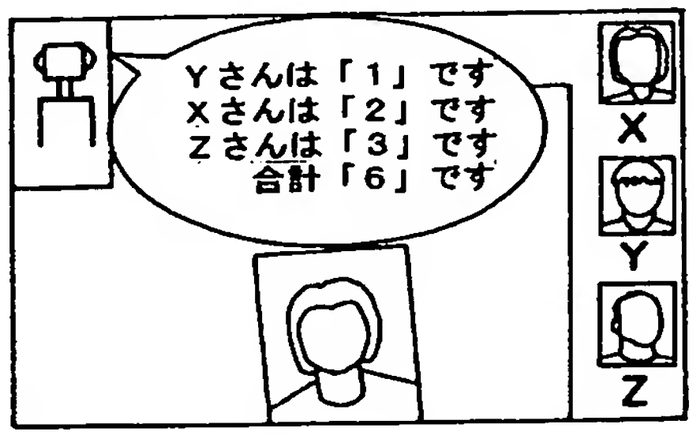
(a)　(b)　(c)



(d)　(e)　(f)

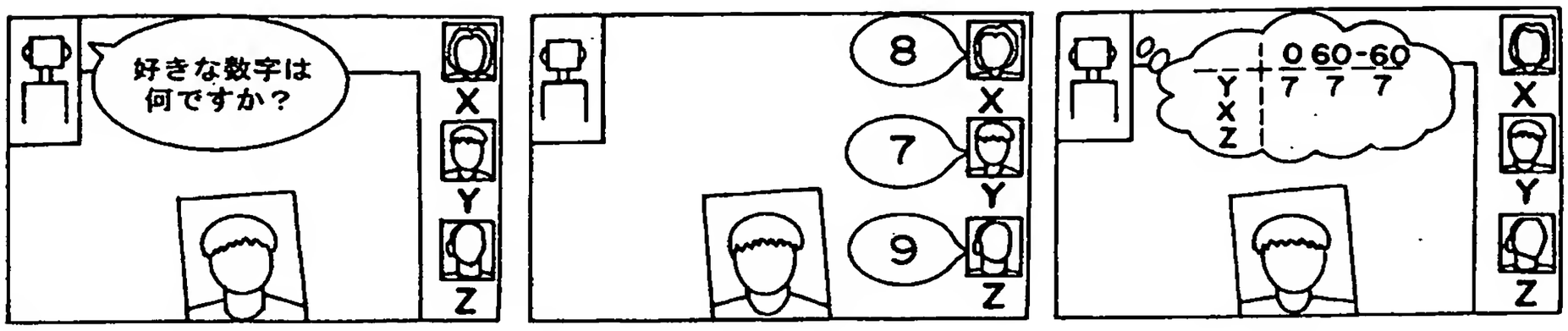


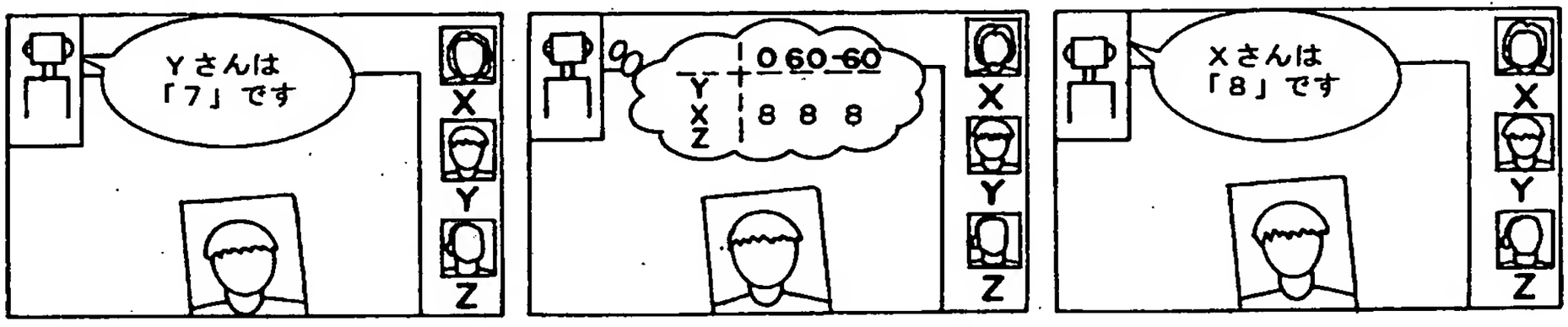
(g)　(h)

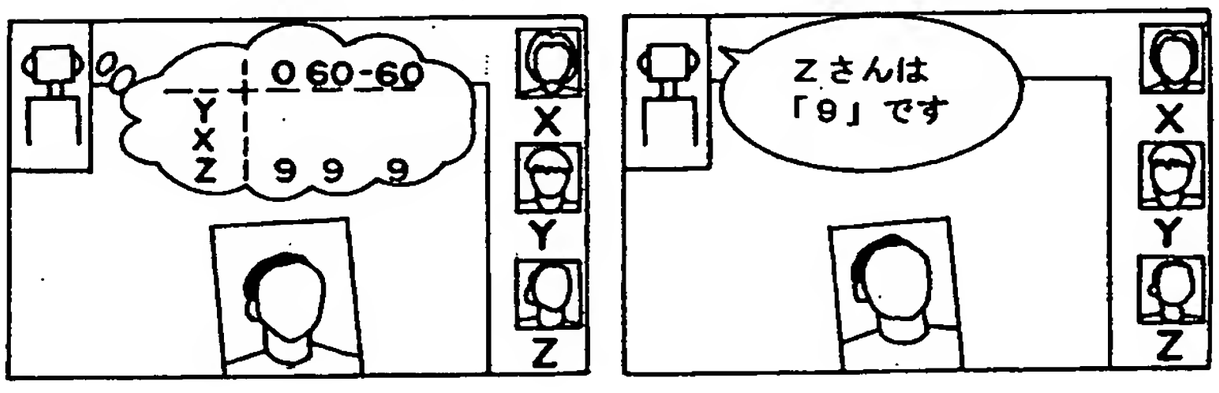


(i)

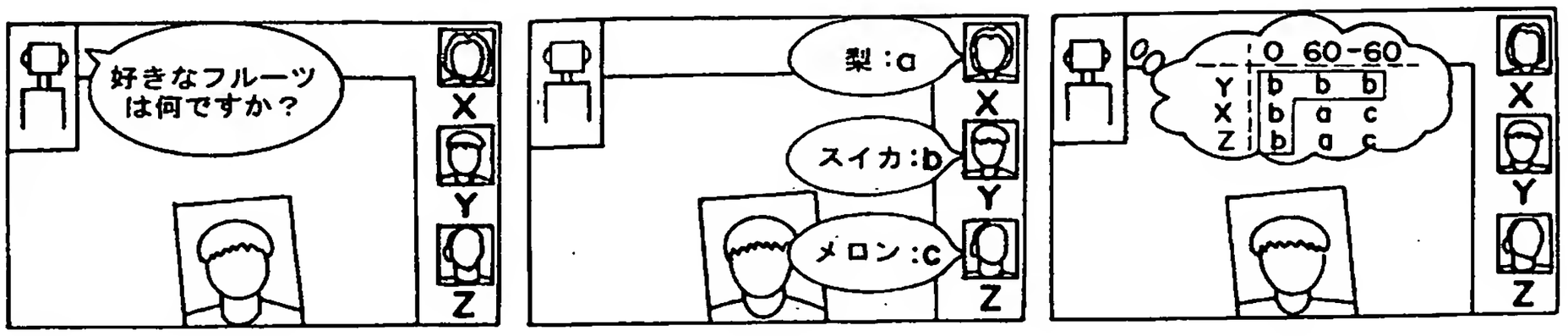
【図１１】


好きな数字は
何ですか？
X
Y
Z
8
7
9
0 60 -60
7 7 7
Y
X
Z
(a)
(b)
(c)

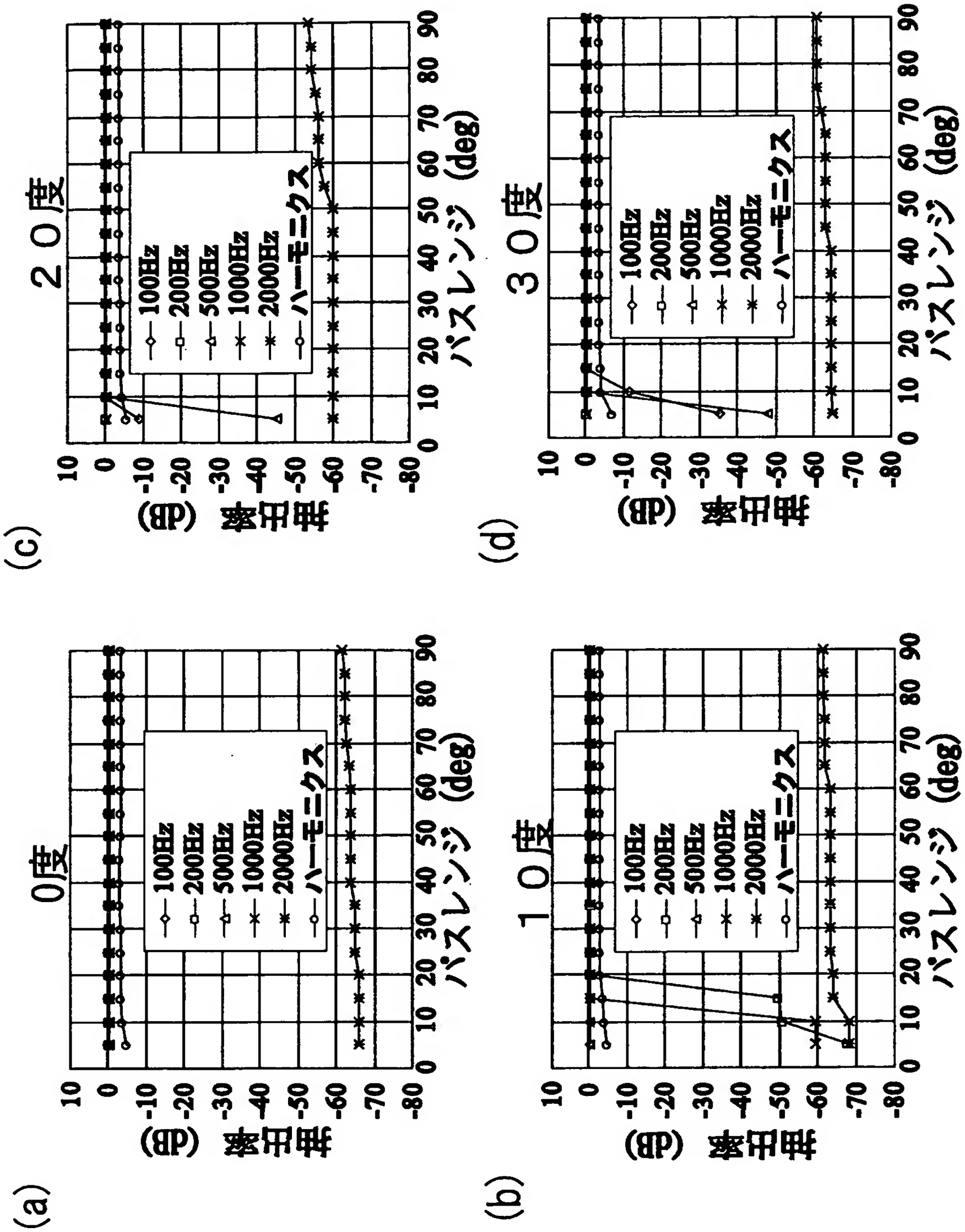

Yさんは
「7」です
0 60 -60
Y
X
Z
8 8 8
Xさんは
「8」です
(d)
(e)
(f)


0 60 -60
Y
X
Z
9 9 9
Zさんは
「9」です
(g)
(h)

【図１２】

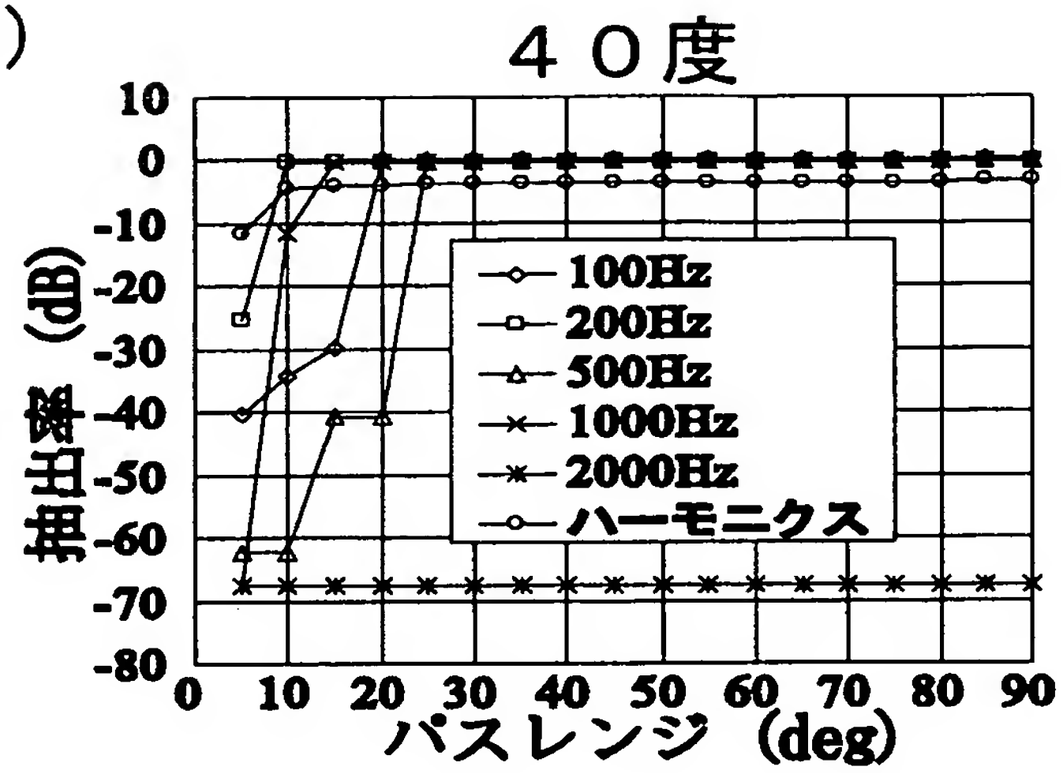

好きなフルーツ
は何ですか？
X
Y
Z
梨：a
スイカ：b
メロン：c
0 60-60
Y b b b
X b a c
Z b a c
(a)
(b)
(c)
(d)
(e)
(f)
Yさんは
「スイカ」です
Xさんは
「梨」です
Zさんは
「メロン」です
(g)
(h)
(i)

【図１３】

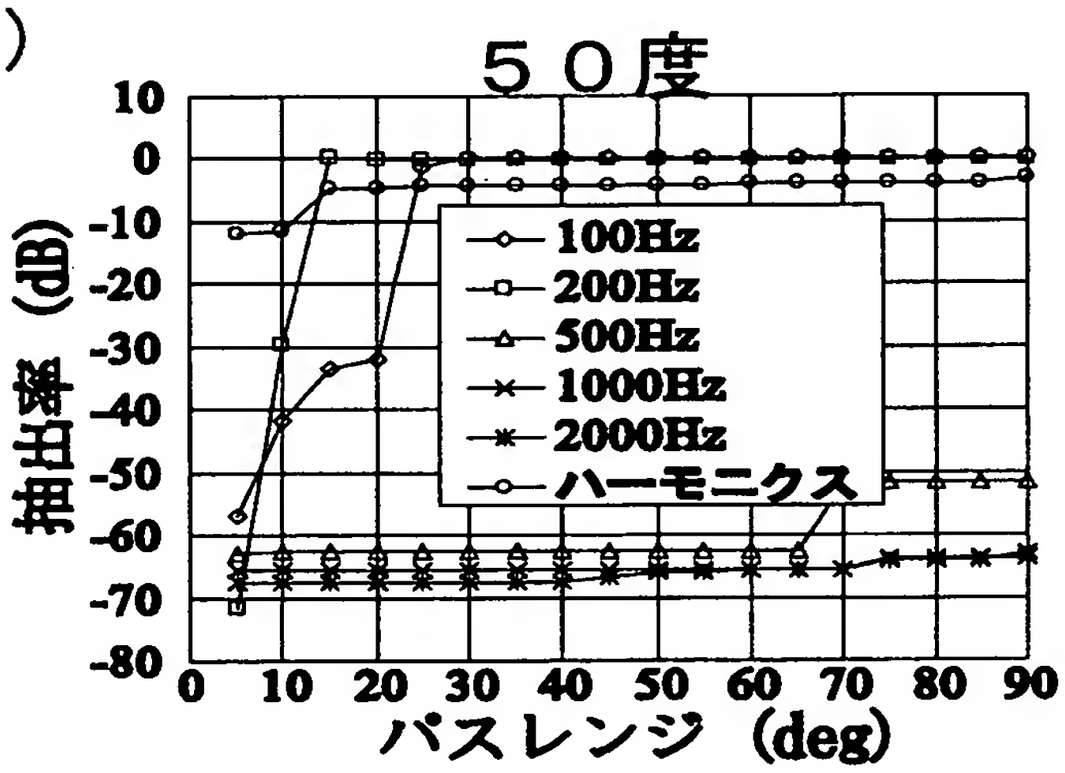

(a)
0度
100Hz
200Hz
500Hz
1000Hz
2000Hz
ハーモニクス
パスレンジ (deg)
0 10 20 30 40 50 60 70 80 90
挿出率 (dB)
10 0 -10 -20 -30 -40 -50 -60 -70 -80
(b)
10度
100Hz
200Hz
500Hz
1000Hz
2000Hz
ハーモニクス
パスレンジ (deg)
0 10 20 30 40 50 60 70 80 90
挿出率 (dB)
10 0 -10 -20 -30 -40 -50 -60 -70 -80
(c)
20度
100Hz
200Hz
500Hz
1000Hz
2000Hz
ハーモニクス
パスレンジ (deg)
0 10 20 30 40 50 60 70 80 90
挿出率 (dB)
10 0 -10 -20 -30 -40 -50 -60 -70 -80
(d)
30度
100Hz
200Hz
500Hz
1000Hz
2000Hz
ハーモニクス
パスレンジ (deg)
0 10 20 30 40 50 60 70 80 90
挿出率 (dB)
10 0 -10 -20 -30 -40 -50 -60 -70 -80

【図１４】

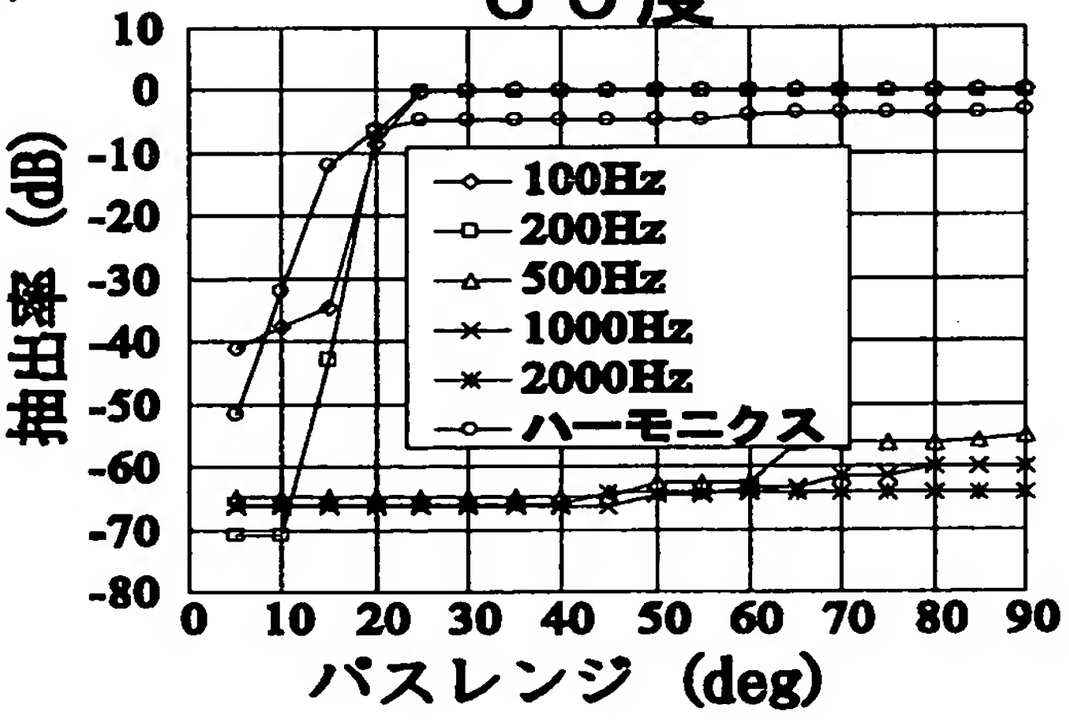
（ａ）


４０度
抽出率（dB）
10
0
-10
-20
-30
-40
-50
-60
-70
-80
0 10 20 30 40 50 60 70 80 90
パスレンジ（deg）
100Hz
200Hz
500Hz
1000Hz
2000Hz
ハーモニクス

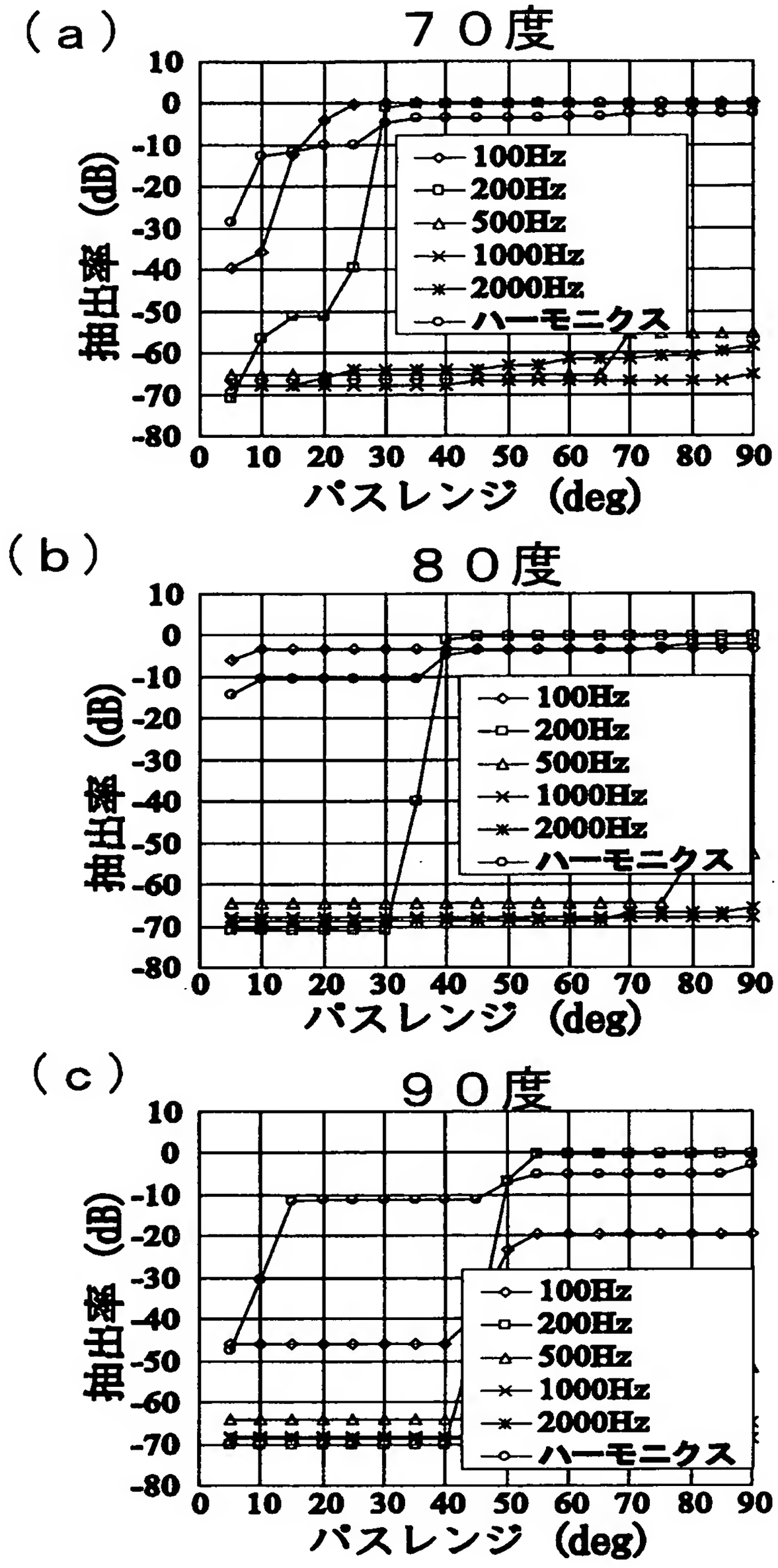
（ｂ）


５０度
抽出率（dB）
10
0
-10
-20
-30
-40
-50
-60
-70
-80
0 10 20 30 40 50 60 70 80 90
パスレンジ（deg）
100Hz
200Hz
500Hz
1000Hz
2000Hz
ハーモニクス

（ｃ）


６０度
抽出率（dB）
10
0
-10
-20
-30
-40
-50
-60
-70
-80
0 10 20 30 40 50 60 70 80 90
パスレンジ（deg）
100Hz
200Hz
500Hz
1000Hz
2000Hz
ハーモニクス

【図１５】


（ａ）
７０度
10
0
-10
-20
-30
-40
-50
-60
-70
-80
出力振幅（dB）
100Hz
200Hz
500Hz
1000Hz
2000Hz
ハーモニクス
0 10 20 30 40 50 60 70 80 90
パスレンジ（deg）
（ｂ）
８０度
10
0
-10
-20
-30
-40
-50
-60
-70
-80
出力振幅（dB）
100Hz
200Hz
500Hz
1000Hz
2000Hz
ハーモニクス
0 10 20 30 40 50 60 70 80 90
パスレンジ（deg）
（ｃ）
９０度
10
0
-10
-20
-30
-40
-50
-60
-70
-80
出力振幅（dB）
100Hz
200Hz
500Hz
1000Hz
2000Hz
ハーモニクス
0 10 20 30 40 50 60 70 80 90
パスレンジ（deg）

【書類名】　要約書

【要約】

【課題】　各音源からの分離された音についての認識を行なうようにしたロボット視聴覚システムを提供する。

【解決手段】　聴覚モジュール２０，顔モジュール３０，ステレオモジュール３７，モータ制御モジュール４０と、これらの各モジュールを制御するアソシエーションモジュール５０とを備え、聴覚モジュールが、複数の音響モデルにより音声認識を行ない、各音響モデルによる音声認識結果をセレクタにより統合して、これらの音声認識結果のうち最も信頼性の高い音声認識結果を判断するように構成されている。

【選択図】　図４

【書類名】　出願人名義変更届（一般承継）
【提出日】　平成15年10月31日
【あて先】　特許庁長官　殿
【事件の表示】
　【出願番号】　特願2002-365764
【承継人】
　【識別番号】　503360115
　【住所又は居所】　埼玉県川口市本町四丁目１番８号
　【氏名又は名称】　独立行政法人科学技術振興機構
　【代表者】　沖村　憲樹
　【連絡先】　〒１０２－８６６６　東京都千代田区四番町５－３　独立行政法人科学技術振興機構　知的財産戦略室　佐々木吉正　ＴＥＬ　０３－５２１４－８４８６　ＦＡＸ　０３－５２１４－８４１７
【提出物件の目録】
　【物件名】　権利の承継を証明する書面　1
　【援用の表示】　平成１５年１０月３１日付提出の特第許３４６９１５６号にかかる一般承継による移転登録申請書に添付のものを援用する。
【物件名】　登記簿謄本　1
　【援用の表示】　平成１５年１０月３１日付提出の特第許３４６９１５６号にかかる一般承継による移転登録申請書に添付のものを援用する。

特願２００２－３６５７６４

出　願　人　履　歴　情　報

識別番号　　［３９６０２０８００］

１．変更年月日　　１９９８年　２月２４日
［変更理由］　　名称変更
住　所　　埼玉県川口市本町４丁目１番８号
氏　名　　科学技術振興事業団

出証番号　出証特２０１１－３０１２９０３

特願２００２－３６５７６４

出　願　人　履　歴　情　報

識別番号　［５０３３６０１１５］

１．変更年月日　２００３年１０月　１日
［変更理由］　新規登録
住　所　埼玉県川口市本町４丁目１番８号
氏　名　独立行政法人　科学技術振興機構

２．変更年月日　２００４年　４月　１日
［変更理由］　名称変更
住　所　埼玉県川口市本町４丁目１番８号
氏　名　独立行政法人科学技術振興機構

３．変更年月日　２００９年　８月２０日
［変更理由］　住所変更
住　所　東京都千代田区四番町５－３　サイエンスプラザ５Ｆ
氏　名　独立行政法人科学技術振興機構

４．変更年月日　２００９年１０月　８日
［変更理由］　住所変更
住　所　埼玉県川口市本町四丁目１番８号
氏　名　独立行政法人科学技術振興機構